THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *17/12/2019* **par :**

MALIK IRAÏN

Plateforme d'analyse de performances des méthodes de localisation des données dans le Cloud basées sur l'apprentissage automatique exploitant des délais de message

---

JURY

SAMIA BOUZEFRANE	Professeure au CNAM Paris	Rapporteure
CONGDUC PHAM	Professeur à l'Université de Pau	Rapporteur
FRANCINE KRIEF	Professeure à l'ENSEIRB-MATMECA	Examinatrice
JEAN-MARC PIERSON	Professeur à l'Université Toulouse 3	Examinateur
ZOUBIR MAMMERI	Professeur à l'Université Toulouse 3	Directeur de thèse
JACQUES JORDA	MCF-HDR à l'Université Toulouse 3	Co-Directeur de thèse

---

École doctorale et spécialité :
MITT : Domaine STIC : Réseaux, Télécoms, Systèmes et Architecture

Unité de Recherche :
Institut de Recherche en Informatique de Toulouse (UMR 5505)

Directeur(s) de Thèse :
Zoubir MAMMERI et *Jacques JORDA*

Rapporteurs :
Samia BOUZEFRANE et *Congduc PHAM*

Remerciements

Avant tout, j'aimerais remercier mes directeurs de thèse : Zoubir Mammeri et Jacques Jorda, ça n'a pas toujours été facile de maintenir une motivation constante durant la durée de cette thèse, mais ils m'ont beaucoup encouragé, surtout lors de la rédaction de ce manuscrit, et j'ai beaucoup appris auprès d'eux.

Faire une thèse, c'est aussi la soutenir devant un jury et je suis très reconnaissant envers les personnes ayant accepté de faire partie du mien : Samia Bouzefrane, Congduc Pham, Francine Krief et Jean-Marc Pierson.

Une page de remerciements sans mentionner les personnes avec qui j'ai partagé mon bureau ne serait pas possible. Je remercie donc ici Ghada et Usman pour m'avoir supporté. J'en profite pour remercier aussi Rahim, pour ses précieux conseils, autant scientifiques que personnels, son soutien et nos longues discussions *géopolitiques* (même le weekend !).

Toujours dans le cadre de la thèse, je veux remercier mes collègues de l'IRIT : Ophélie, Tristan, Thi ainsi que les personnes occupant ou ayant occupé le bureau 467 : Tanissia, Léo, Chaopeng, Gustavo et Zongyi. j'ai passé les 3 ans (et un peu plus) de ma thèse avec de vous. Sans les (innombrables) pauses repas et café, les sorties et les barbecues avec vous, cette *aventure* n'aurait pas été aussi mémorable.

Étant donné leur soutien et leurs oreilles attentives à mes *complaintes*, je remercie énormément Zoé et Clem. Je ne sais pas si vous lirez ces mots, mais j'en profite quand même pour vous féliciter pour votre *projet*, que vous avez su monter en donnant énormément de votre temps, sans savoir si des personnes seraient intéressées pour le reprendre... Et ça a fonctionné ! Bravo !

Enfin, je remercie ma famille pour m'avoir toujours encouragé à faire des études qui me passionnent et à continuer le plus loin possible. Je les remercie d'être toujours là pour moi, quoi qu'il arrive et à n'importe quel moment de la journée. D'abord ma mère, Yamina, puis mes sœurs et mon frère, Sonia et Nadia et Mehdi. Je n'oublie pas mes nièces et mon neveu, Lily, Julia, Hidaya et Tom, qui me forcent à prendre de vraies vacances et ne me laissent pas le temps de penser à autre chose durant les trop rares moments pendant lesquels on se voit.

Abstract

Cloud usage is a necessity today, as data produced and used by all types of users (individuals, companies, administrative structures) has become too large to be stored otherwise. It requires to sign, explicitly or not, a contract with a cloud storage provider. This contract specifies the levels of quality of service required for various criteria. Among these criteria is the location of the data.

However, this criterion is not easily verifiable by a user. This is why research in the field of data localization verification has led to several studies in recent years, but the proposed solutions can still be improved. The work proposed in this thesis consists in studying solutions of location verification by a user, i.e. solutions that estimate data location and operate using landmarks. The implemented approach can be summarized as follows : exploiting communication delays and using network time models to estimate, with some distance error, data location. To this end, the work carried out is as follows :

— A survey of the state of the art on the different methods used to provide users with location information.

— The design of a unified notation for the methods studied in the survey, with a proposal of two scores to assess methods.

— Implementation of a network measurements collecting platform. Thanks to this platform, two datasets were collected, at both national level and international level. These two data sets are used to evaluate the different methods presented in the state of the art survey.

— Implementation of an evaluation architecture based on the two data sets and the defined scores. This allows us to establish the quality of the methods (success rate) and the quality of the results (accuracy of the result) thanks to the proposed scores.

Résumé

L'utilisation du cloud est une nécessité aujourd'hui, les données produites et utilisées par tous les types d'utilisateurs (individus particuliers, entreprises, structures administratives) ayant atteint une masse trop importante pour être stockées autrement. L'utilisation du cloud nécessite la signature, explicite ou non, d'un contrat avec un fournisseur de service de stockage. Ce contrat mentionne les niveaux de qualité de service requis selon différents critères. Parmi ces critères se trouve la localisation des données.

Cependant, ce critère n'est pas facilement vérifiable par un utilisateur. C'est pour cela que la recherche dans le domaine de la vérification de localisation de données a suscité plusieurs travaux depuis quelques années, mais les solutions proposées restent encore perfectibles. Le travail proposé dans le cadre de cette thèse consiste à étudier les solutions de vérification de localisation par les clients, c'est-à-dire les solutions estimant la localisation des données et fonctionnant à l'aide de points de repère. L'approche à investiguer peut être résumée comme suit : en exploitant les délais de communication et en utilisant des modèles de temps de traversée du réseau, estimer, avec une certaine erreur de distance, la localisation des données. Pour cela, le travail réalisé est le suivant :

— Une revue de l'état de l'art des différentes méthodes permettant aux utilisateurs de connaitre la localisation de leurs données.

— La conception d'une notation unifiée pour les méthodes étudiées dans la revue de l'état de l'art, avec une proposition de deux scores pour évaluer et comparer les méthodes.

— La mise en place d'une plateforme de collecte de mesures réseau. Grâce à cette plateforme, deux jeux de données ont été récoltés, un au niveau national et l'autre un niveau mondial. Ces deux jeux de données permettent d'évaluer les différentes méthodes présentées dans la revue de l'état de l'art.

— La mise en place d'une architecture d'évaluation à partir des deux jeux de données et des scores définis, afin d'établir la qualité des méthodes (taux de succès) et la qualité des résultats (précision du résultat) grâce aux scores proposés.

Table des matières

Table des figures

Liste des tableaux

Introduction

1 Contexte

De nos jours, les masses de données manipulées par les individus, les entreprises ou les structures administratives sont devenues tellement importantes que leur stockage sur le cloud devient une nécessité. Aussi, les réseaux de communication se multiplient et deviennent accessibles à tous, quasiment partout et à n'importe quel moment. Ainsi l'utilisateur peut accéder à ses données stockées dans le cloud de manière quasi-instantanée et transparente. Les utilisateurs signent des contrats, dits des contrats de qualité de service, avec les fournisseurs de service de stockage dans lesquels sont mentionnés les niveaux de qualité de service requis par ces utilisateurs selon différents critères. Ces critères peuvent être : le temps de réponse, la disponibilité du service, la sécurité ou bien la localisation des données, parmi d'autres.

Comme les données critiques, en ce qui concerne la vie privée des utilisateurs ou les secrets des institutions, des états ou des entreprises doivent être stockées à des endroits considérés comme dignes de confiance, les fournisseurs de service de stockage doivent garantir et prouver que les données sont stockées dans la ville, le pays ou le continent du propriétaire des données. Des lois récentes en Europe, en Australie, en Chine, et dans plusieurs autres pays imposent aux fournisseurs de service de stockage de garantir que les données ne quittent jamais la zone géographique où elles sont censées être stockées.

La garantie de localisation se base sur différentes fonctions et différentes hypothèses de confiance dans la chaîne de communication entre les clients et les serveurs. La robustesse du mécanisme de localisation dépend du prix que les utilisateurs sont capables de supporter. Si l'utilisateur n'a pas totalement confiance envers son fournisseur, deux approches peuvent être utilisées pour garantir la localisation des données : utilisation d'un tiers de confiance ou vérification par le client.

La recherche dans le domaine de la vérification de localisation de données a suscité plusieurs travaux depuis quelques années et les solutions proposées restent encore perfectibles, notamment en ce qui concerne la qualité des résultats et le coût pour l'utilisateur. La comparaison de ces solutions est rendue difficile par les différents indicateurs proposés par leurs auteurs, chacune se valorisant dans un contexte expérimental différent.

2 Contributions

Le travail proposé dans le cadre de cette thèse consiste à étudier les solutions de vérification de localisation par les clients, c'est-à-dire les solutions estimant la localisation des données et fonctionnant à l'aide de points de repère. L'approche à investiguer peut être résumée comme suit : en exploitant les délais de communication (lors des requêtes d'accès aux données) et en utilisant des modèles de temps de traversée du réseau, estimer, avec une certaine erreur de distance, la localisation des données. Pour cela, le travail réalisé est le suivant :

— Une revue de l'état de l'art des différentes méthodes permettant aux utilisateurs d'obtenir l'information de localisation. Cette revue propose une classification des différentes méthodes de vérification par l'utilisateur et met en évidence les limites de ces solutions.

— La conception d'un modèle unifié pour les méthodes étudiées dans la revue de l'état de l'art. Ce modèle permet d'homogénéiser les notations et de proposer des scores adaptés à l'ensemble des méthodes. Deux scores sont proposés.

— La mise en place d'une plateforme de collecte de mesures réseau. Grâce à cette plateforme, deux jeux de données ont été récoltés, un au niveau national et l'autre un niveau mondial. Le jeu de données national a été collecté grâce à la plateforme Grid5000 et celui mondial grâce à Amazon Web Services. Ces deux jeux de données permettent d'évaluer les différentes méthodes présentées dans la revue de l'état de l'art.

— La mise en place d'une architecture d'évaluation à partir des deux jeux de données et des scores définis, afin d'établir la qualité des méthodes (taux de succès) et la qualité des résultats (précision du résultat). Deux évaluations sont réalisées, en fonction des deux jeux de données : une au niveau national et l'autre au niveau mondial.

3 Organisation du manuscript

Le manuscrit est organisé en deux parties, divisées de la façon suivante :

— La partie I présente une revue et une analyse de l'état de l'art du sujet. Elle est composée de trois chapitres :

  — Le chapitre 1 détaille les motivations et les raisons pour lesquelles les utilisateurs souhaitent connaître et choisir la localisation des données stockées dans le Cloud. D'une part, des raisons de sécurité et de performances guident à choisir la position de leurs données. D'autres part ce sont des raisons de respect de la législation qui poussent les utilisateurs à choisir et connaître la localisation de leurs données.

  — Le chapitre 2 présente le premier type de méthodes permettant la localisation des données dans le Cloud, c'est-à-dire les méthodes utilisant un tiers de confiance et nécessitant du matériel ou du logiciel spécifique au sein de l'infrastructure du fournisseur pour assurer la localisation. Les limites de ces méthodes sont aussi abordées.

  — Le chapitre 3 présente et propose une classification du deuxième type de méthodes permettant la localisation des données dans le Cloud, c'est-à-dire les méthodes d'estimation de la localisation dans le Cloud à l'aide de points de repère. Ces méthodes ne nécessitent aucune action du fournisseur et leur mise en place et fonctionnement sont assurés par l'utilisateur souhaitant vérifier la localisation des données. Comme pour le chapitre précédent, les limites de ces méthodes sont présentées. Ce sont ces méthodes qui seront étudiées dans les contributions.

— La Partie II présente les contributions réalisées au cours de cette thèse. Elle est composée de trois chapitres :

  — Le chapitre 4 propose un cadre pour généraliser les notations, le déroulement des méthodes étudiées. Le déroulement est découpé en phases d'apprentissage, de vérification et d'évaluation. Deux scores sont proposés pour la phase d'évaluation.

  — Le chapitre 5 décrit la plateforme de collecte des deux jeux de données. Ces deux jeux de données ont été collectés respectivement au niveau national et au niveau mondial. Il présente les caractéristiques

générales des jeux de données, telles que le nombre d'échantillons, les RTTs moyens entre deux nœuds, etc. Le nettoyage réalisé sur ces jeux de données est aussi abordé.

— Le chapitre 6 présente l'architecture mise en place pour évaluer les différentes méthodes à l'aide des mesures (délais de messages) collectées. Les résultats obtenus au niveau national et mondial sont présentés et commentés pour proposer un aspect méthodologique pour l'utilisation des approches.

Le manuscrit termine par une conclusion, synthétisant les contributions de cette thèse et apportant les perspectives futures de ces travaux.

Première partie

Revue de l'état de l'art

Chapitre 1

Position des données stockées dans le Cloud

Sommaire

1.1 Introduction

Le Cloud computing est un modèle qui permet de mettre à disposition de ses utilisateurs différents services tout en le déchargeant le plus possible de la complexité, de l'administration et de la maintenance liées à ces services. Il est question de modèles as-a-service, qualifiant ainsi les différentes solutions proposées par les fournisseurs de service Cloud [1].

Le stockage de données fait partie des solutions proposées en tant que service Cloud. Pour un utilisateur, il se différencie d'un stockage traditionnel par trois points principaux [2] :

— La mise à disposition de protocoles permettant un accès facilité aux données à travers Internet. Par exemple, la plupart des fournisseurs de stockage Cloud permettent l'accès aux données par HTTP(S) au sein d'un navigateur ou en utilisant une API REST.

— L'élasticité du stockage. Si plus d'espace de stockage est nécessaire, il peut être automatiquement alloué à l'utilisateur. De même, si moins d'espace est nécessaire, moins d'espace sera alloué à l'utilisateur et il ne sera facturé que pour l'espace réellement utilisé.

— La mise en place d'une redondance des données est facilitée, de telle sorte qu'une défaillance chez le fournisseur ne mette pas en danger l'existence et l'accessibilité aux données. Des copies des données peuvent être réparties dans différents lieux géographiques pour pouvoir assurer une continuité d'accès même en cas de problèmes dans l'un des centres de stockage.

De plus, même si l'abstraction fournie par le Cloud le masque à l'utilisateur, les données sont bien physiquement stockées dans un, voire plusieurs, lieu de stockage donné. C'est donc de la zone géographique des lieux de stockage des données dont il est question lorsque la localisation des données est abordée. Cette zone géographique peut être définie de différentes manières. Par ordre croissant de précision et de façon non exhaustive, elle peut être donnée comme :

— Les coordonnées GPS du lieu de stockage.

— La ville dans laquelle se situe le lieu de stockage.

— Le pays, l'état ou la région dans laquelle se situe le lieu de stockage.

Dans ce qui suit, nous donnons un aperçu de l'importance de la position géographique des données stockées dans le Cloud en particulier pour des raisons de QoS, de sécurité et de performance, et de respect des lois.

1.2 Gestion de la QoS dans le Cloud

1.2.1 Contrats de SLA

Les contrats de service Cloud sont régis par des ententes de niveau de service (SLA ou Service-Level Agreement). Le SLA permet à l'utilisateur et au fournisseur de se mettre d'accord sur le niveau de service attendu et les pénalités en cas de non-respect du contrat. Le SLA contient alors différentes clauses de qualité de service (QoS ou Quality of Service), qui concernent généralement la disponibilité du service et la durabilité du stockage, ainsi que les indicateurs permettant le contrôle de la qualité des différentes clauses de QoS [3].

Par exemple, une clause apparaissant dans la totalité des contrats SLA grand public des fournisseurs de service Cloud est le pourcentage de disponibilité du service par l'utilisateur à travers Internet [4]. Si l'accessibilité au service, pour la période de temps définie par le SLA, est moindre que le pourcentage défini par le SLA, une pénalité financière, généralement en termes de crédits pour le service donné, est versée par le fournisseur au client. Il est aussi possible de retrouver régulièrement des clauses concernant le temps de réponse, qui correspond au temps entre la requête d'accès au service et la réponse reçue, ainsi que la fiabilité, qui défini le nombre d'erreurs autorisées sur une période donnée. Quand bien même la confirmation du non-respect de la clause de QoS n'est possible que si le fournisseur de service est honnête, car il est celui qui a les moyens de vérifier ses infrastructures pour confirmer ou infirmer un problème, il n'est pas dans son intérêt de frauder, le SLA étant un document ayant une valeur légale.

Réciproquement, une pénalité financière peut être attribuée à l'utilisateur [5]. En effet, le SLA régissant les engagements des deux parties, fournisseur et utilisateur, certains contrats imposent une pénalité financière à l'utilisateur pour toute fausse déclaration de non-respect de QoS.

1.2.2 La localisation des données : une clause de QoS

La zone géographique de stockage des données, ou localisation des données, dans le Cloud est un critère de QoS. En effet, ce critère peut se retrouver au sein des SLA sous deux formes :

— La duplication des lieux de stockage, afin de réaliser une redondance des données, pour des raisons de sécurité et de tolérance aux fautes ou bien pour des raisons de performances avec de la géo-réplication.

— La précision d'une zone géographique pour le ou les lieux de stockage, afin

de garantir leur présence et leur accès au sein d'une limite administrative. Et ce, principalement pour des raisons légales et de respect de la vie privée.

Les deux formes ne sont cependant pas exhaustives. En effet un utilisateur peut vouloir une duplication des lieux de stockage, mais tous limités à une certaine zone géographique, par exemple à l'intérieur d'un pays.

De plus, il faut noter que contrairement aux autres critères de QoS habituellement présents dans les SLA, la localisation des données n'est pas facilement vérifiable par un utilisateur, dans le contexte Cloud actuel. En effet, il est facile de vérifier le pourcentage de disponibilité du service, le temps d'accès ou la fiabilité. Pour chacun de ces trois critères, la mesure est à la portée des utilisateurs. Il s'agit respectivement de littéralement compter la durée d'indisponibilité, le temps que met une requête à être exécutée et le nombre d'erreurs sur une période donnée. Ensuite l'utilisateur concerné n'a plus qu'à comparer ces mesures avec celles fournies dans le contrat. Pour la localisation des données, il est impossible de « simplement » la mesurer et il faut mettre en place des mécanismes un peu plus complexes afin de s'en assurer.

Même si un utilisateur grand public, consommateur de logiciels en tant que service (SaaS ou Software-as-a-Service), ne voit pas toujours d'intérêt à pouvoir spécifier et connaître la localisation de ses données, différentes raisons peuvent pousser un utilisateur à limiter les zones géographiques dans lesquelles seront stockées ses données, et ainsi d'avoir besoin de s'assurer du respect de la clause de QoS correspondante. Ces raisons sont d'autant plus importantes si l'utilisateur est une entreprise ou une entité publique. D'une part ces raisons sont plutôt techniques, comme assurer la sécurité des données ainsi qu'un certain niveau de performance sur l'accès aux données. D'autre part, ce sont des raisons de respect de la législation et de la vie privée.

1.3 Sécurité et performances

Dupliquer les données ou préciser une zone géographique sur les lieux de stockage permet de garantir un certain niveau de sécurité et de performances pour l'utilisateur. Dans cette section, il sera montré comment le respect des clauses de QoS concernant la localisation des données permet d'assurer ce niveau.

1.3.1 Sécurité

En termes de sécurité, trois raisons ont été identifiées pour justifier le besoin de connaitre la localisation des données.

Protéger les données en cas d'accident

Les données étant stockées sur des supports physiques à l'intérieur de centres de données, elles ne sont pas à l'abri de catastrophes naturelles, telles que les tornades, tsunamis et éruptions volcaniques, ni d'accidents moins « extraordinaires » comme les incendies ou inondations. Ces accidents peuvent avoir pour conséquence la perte des données.

En effet, les supports de stockages peuvent être directement endommagés par l'accident, comme en juin 2011 où une tornade a détruit l'hôpital « St. John's Regional Medical Center » dans le Missouri ainsi que le centre de données adjacent qui contenait des données médicales de patients, issues de l'hôpital [6].

Savoir où sont stockées les données permet donc d'estimer et limiter les risques de les perdre suite à un accident naturel, même si tout n'est pas prévisible [7].

Empêcher l'accès aux données à une entité extérieure

Connaître et choisir le lieu de stockage de ses données permet de s'assurer qu'une entité extérieure à l'utilisateur et au fournisseur de service, par exemple un organisme gouvernemental du pays dans lequel se trouvent les données, n'accède ni n'intercepte les données. En effet, prenons l'exemple de données françaises qui se retrouvent stockées en Russie parce qu'un utilisateur a voulu externaliser le stockage de ses données. D'après le principe de souveraineté des données le gouvernement russe a potentiellement le droit d'accéder à ces données (figure 1.1) car elles sont soumises aux lois du territoire sur lesquelles elles sont stockées, dans ce cas la Russie.

Si les données stockées ne sont que des données personnelles d'utilisateurs grand public le problème peut être considéré comme négligeable, mais il est d'autant plus important quand les données le sont aussi. Il ne serait pas souhaitable que des données sensible du gouvernement français ou de grandes industries soient accessibles à des gouvernements étrangers.

Il est possible d'observer ce genre d'accès par une entité extérieure. En effet, aux États-Unis, grâce aux révélations d'Edward Snowden sur la surveillance de la NSA, l'accès à des données internes à Google et Yahoo ! par l'agence de sécurité grâce au projet MUSCULAR a été exposé [8].

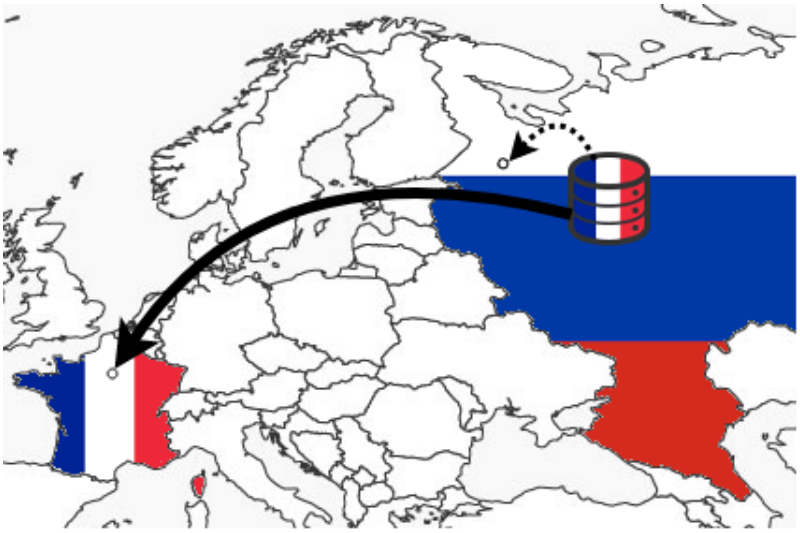


Figure 1.1 – Accès Russe à des données françaises

Il est aussi à noter que plus récemment, le CLOUD Act [9] remet en question la souveraineté des données. Il ne se fonde pas sur le lieu de stockage des données mais sur la « nationalité » du fournisseur, en considérant les données comme soumises aux lois états-uniennes même en cas d'extraterritorialité.

Garantir un accès aux données au sein d'une zone administrative

En stockant des données au sein d'une zone administrative, comme un pays, il est possible de s'assurer que le service soit accessibles aux utilisateurs situés dans cette zone administrative. Effectivement, si les données sont stockées en dehors de la zone administrative des utilisateurs, c'est-à-dire dans un autre pays, de nouvelles lois parfois issues de tensions géopolitiques peuvent empêcher les utilisateurs extérieurs à la zone administration dans laquelle sont stockées les données d'y accéder. Comme illustré par la figure 1.2, c'est ce que la Russie tente de mettre en place en se coupant de l'Internet mondial, résultant en des données inaccessibles pour des utilisateurs français si elles y sont stockées.

Le blocage n'est pas forcément le fait du gouvernement du pays contenant les données mais peut aussi venir du pays de l'utilisateur. Même si ce ne sont pas spécifiquement des services de stockages de données, il est déjà possible d'observer ce genre de blocages. C'est le cas en Chine où certains sites étrangers comme la plupart des services Google sont bloqués [10]. Ici ce sont les autorités chinoises qui bloquent l'accès à des services extérieurs depuis l'intérieur de la Chine. Un autre exemple est certains sites, principalement des journaux

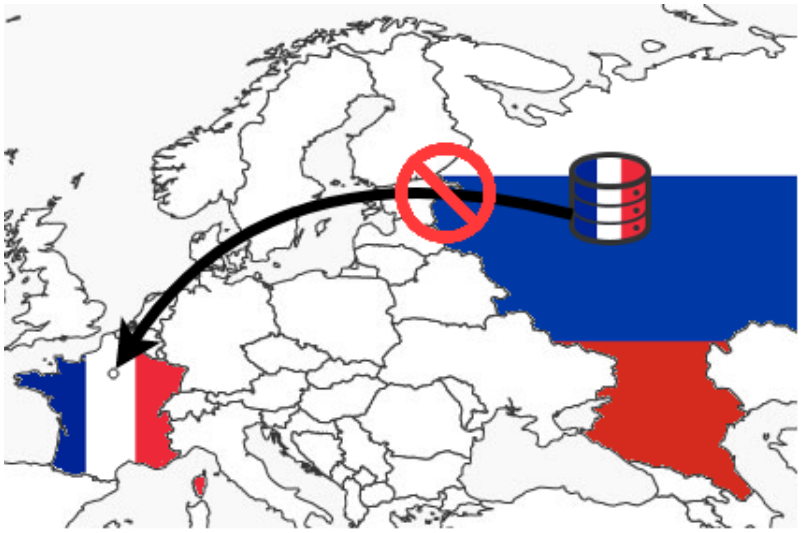


Figure 1.2 – Données françaises non accessibles en France

américains, bloquant l'accès depuis l'Union Européenne, parce que ces sites sont non conformes au nouveau règlement RGPD. Dans ce cas, le blocage ne vient pas d'un gouvernement mais d'une entité privée.

1.3.2 Performances pour l'accès utilisateur

Les améliorations de performance pour l'utilisateur se font principalement ressentir lorsque les données sont répliquées sur différents sites. C'est le principe de fonctionnement des Content Delivery Network (CDN ou réseau de diffusion de contenu) [11] et qui permettent à des améliorations pour l'utilisateur sur les deux points suivants.

Réduction du délai

En choisissant une zone géographique proche des accès aux données, autrement dit en plaçant les données proches de l'utilisateur, le délai d'accès est amélioré. Si les accès au service sont limités à une zone géographique précise il est intéressant d'essayer d'y placer les données afin qu'elles se situent au plus proche de l'utilisateur. De plus, si les accès ne sont pas limités à une seule zone géographique, il est toujours possible de choisir plusieurs zones géographiques de stockage avec réplication des données pour obtenir un effet similaire.

Continuité d'accès au service

Différents problèmes sur le réseau peuvent empêcher le maintien d'un accès continu au service, diminuant ainsi pour l'utilisateur le taux de disponibilité du service. Un panne de réseau au sein d'un des lieux de stockage empêche les services hébergés en ce lieu d'être accessible depuis le réseau public et donc par l'utilisateur. De plus, la même situation peut apparaître lors d'une attaque de type Distributed Denial of Service (DDoS ou attaque par déni de service). Une des solutions pour mitiger les conséquences de ces problèmes sur l'utilisateur est la duplication des lieux de stockage.

En effet, si un lieu de stockage connait une panne, un autre peut être rapidement configuré pour prendre le relai, le temps pour le fournisseur de corriger le problème.

1.4 Respect de la législation

1.4.1 Cas des données de santés en France

D'après un rapport de Dell EMC, un acteur important dans les systèmes de stockage, et de l'International Data Corporation (IDC) [12], la quantité de données de santé en 2013 était de 153 exaoctets. Leurs prévisions s'accordent à dire qu'en 2020, la quantité de données de santé atteindra 2 314 exaoctets. Les données de santé représentent donc une quantité considérable de données à stocker, ce qui explique que les acteurs produisant ces données, comme les hôpitaux ou les centres de santé, sous-traitent le stockage.

Cependant, les données de santé peuvent contenir des informations confidentielles, telles que les dossiers médicaux des patients, avec l'ensemble des images médicales réalisées, les opérations subies, les traitements suivis et les pathologies ou affections dont souffre le patient. Il n'est donc pas souhaitable qu'elles soient stockées n'importe où, pour des raisons de sécurité et de vie privée, et que les acteurs produisant ces données aient le choix du lieu de stockage.

C'est pour cela qu'en France l'article L.1111-8 [13] du code de la santé publique ainsi que le décret n° 2018-137 du 26 février 2018 relatif à l'hébergement de données de santé à caractère personnel [14] définissent et encadrent l'hébergement des données de santé. Ce décret encadre le stockage des données de santé en demandant à ce qu'un centre de données soit certifié (anciennement agréé) « Hébergeur de Données de Santé » (HDS). Pour effectuer une demande d'agrément, les prérequis sont listés dans le référentiel HDS et plusieurs normes et exigences sont à respecter :

— ISO 27001 « système de gestion de la sécurité des systèmes d'information ».

— ISO 20000 « système de gestion de la qualité des services » (partiellement).

— ISO 27018 « protection des données à caractère personnel » (partiellement).

— Exigences complémentaires aux normes ISO 27001, ISO 20000.

— Exigences relatives à la protection des données de santé à caractère personnel.

— Exigences spécifiques au domaine de la santé.

Les prérequis explicitent aussi que la localisation des données est une exigence à respecter. L'hébergeur doit lister l'ensemble des pays dans lequel les données peuvent être stockés, c'est-à-dire l'ensemble des pays pour lesquels un centre de données est certifié (ou agrée) HDS. De plus l'hébergeur doit permettre à l'utilisateur de choisir le ou les lieux de stockage de ses données, tout en s'assurant que le choix de l'utilisateur pour la localisation est bien respecté.

Le respect des normes et exigences, et donc le respect de celles qui concernent la localisation, est vérifié par un audit avant l'obtention de la certification, valable 3 ans. De plus un audit de contrôle annuel a aussi lieu.

1.4.2 Le RGPD au sein de l'Union Européenne

Le règlement n° 2016/679 aussi connu comme le Règlement Général sur la Protection des Données (RGPD) [15] est un règlement issu de l'Union Européenne (UE) et s'appliquant au sein des 28 États membres. Le RGPD remplace la directive 95/46/CE sur la protection des données personnelles pour les 28 États membres.

Ce règlement a pour but d'apporter un cadre harmonisé relatif à la protection des données personnelles concernant les résidents de l'UE. Les données personnelles, sont définies d'après le RGPD comme celles permettant l'identification d'une personne physique, c'est-à-dire une référence à :

— Un nom.

— Un numéro d'identification.

— Des données de localisation.

— Un identifiant en ligne.

De plus le RGPD applique aussi des dispositions particulières aux données sensibles d'une personne, telles que :

— La prétendue origine raciale ou ethnique.

— Les opinions politiques.

— Les convictions religieuses ou philosophiques.

— L'appartenance syndicale.

— Les données génétiques.

— Les données biométriques servant à l'identification unique.

— Les données de santé.

— la vie sexuelle ou l'orientation sexuelle.

Il s'applique lorsqu'une organisation, de n'importe quel type, traite des données personnelles concernant un résident de l'UE, et ce quelque même si l'organisation est hors de l'UE. Le RGPD a donc une application extraterritoriale et il touche un grand nombre d'acteurs du fait de sa définition large des données personnelles et de son application à n'importe quel type d'acteur les traitant.

Imposer des contraintes sur la position des données, comme imposer aux acteurs de stocker leurs données dans l'UE, n'est pas le but du RGPD. Si des contraintes fortes sur la localisation des données sont nécessaires c'est à la législation des États membres de le prendre en compte. Cependant, la localisation reste quand même importante, car les données personnelles stockées au sein de l'UE et celles stockées hors de l'UE sont soumises à des règles différentes. En effet, les données stockées au sein de l'UE sont considérées comme respectant par défaut le RGPD. Pour pouvoir les transférer en dehors de l'UE il faut qu'une des trois conditions suivante soit respectée :

— Le pays est reconnu par l'UE comme assurant un niveau de protection adéquat : Andorre, Argentine, Guernesey, Israël, Île de Man, Îles Féroé, Japon, Jersey, Nouvelle-Zélande, Suisse, Uruguay, Canada (organisations commerciales) et États-Unis d'Amérique (organisations appartenant au Privacy Shield).

— Des règles d'entreprise contraignantes (BCR) qui assurent un niveau de protection adéquat et qui sont juridiquement contraignantes internationalement.

— Par exception grâce à une autorisation, sous conditions, de la Commission Nationale de l'Informatique et des Libertés (CNIL) en France et

organisme équivalent dans les autres pays de l'UE.

L'aide à la conformité et le contrôle du respect du RGPD en France est effectué par la CNIL qui agit principalement en fonction des plaintes reçues par les utilisateurs. Les organismes équivalent des autres États membres effectuent les même missions.

1.4.3 Ailleurs dans le monde

Le problème de la localisation des données est un problème qui concerne tous les pays, chacun essayant de protéger les données personnelles de ses citoyens. Il n'est donc pas étonnant de voir des législations similaires voire plus forte à celles applicables en France dans d'autres pays au sujet de la localisation des données.

On peut notamment citer les trois pays suivants, hors UE, qui se distinguent par leurs législations plus contraignantes qu'en France.

Australie

En Australie, les données de santé des citoyens ne peuvent être stockés ni traitées en dehors de l'Australie, d'après le My Health Records Act [16]. L'Toutefois, si les données ne contiennent pas d'information permettant l'identification des personnes, elles peuvent être transférées, stockées et manipulées à l'extérieur du territoire australien.

Russie

En Russie, d'après la Loi Fédérale n° 242 [17] sur les procédures de traitement des données personnelles, les données personnelles des citoyens Russes doivent être stockés et manipulées en Russie. Il n'y a aucune exception, les données personnelles ne peuvent pas, d'après la loi, quitter la Russie. De plus le lieu de stockage des données doit être précisé à l'utilisateur dont les données ont été recueillies.

Chine

En Chine, des contraintes de localisation existent sur plusieurs types de données :

— Données financières : d'après une annonce de la Banque Populaire de Chine [18], devant être suivie par les banques en Chine, la collecte, le

traitement et le stockage des données financières personnelles doit se faire sur le territoire chinois. La définition des données financières personnelles étant très large, on peut considérer que l'ensemble des données financières doit être stocké en Chine.

— Données personnelles : en accord avec les lignes directrices du gouvernement chinois [19], les données personnelles doivent être stockées en Chine, sauf consentement explicite de l'utilisateur.

— Données commerciales : conformément à la loi sur les secrets d'États [20], les données commerciales qui constituent des secrets d'État ne peuvent pas être transférées hors de Chine.

Chapitre 2

Garanties de localisation des données dans le Cloud par un tiers de confiance

Sommaire

2.1 Introduction

Comme vu dans le chapitre 1, la localisation des données est une information importante pour un utilisateur. De plus, cette information de localisation n'est pas facilement accessible à l'utilisateur surtout dans un contexte Cloud, contrairement aux autres critères de QoS.

Le paradigme est de décharger l'utilisateur de cette mise en place en lui mettant à disposition un ou plusieurs espaces de stockage, l'utilisateur devant faire confiance à son fournisseur pour le placement de ses données au sein d'une zone géographique précise.

Bien sur, des certifications ou agréments existent pour certains types de données, notamment les données de santé, et ce sont des méthodes qui assurent la localisation des données. Cependant ces certifications ou agréments ne sont pas assez génériques, elles ne prennent pas en compte tous les types de données et ne sont mis en place que dans certains pays - ceux ayant les moyens de garantir leur validité. De plus, elles ne permettent pas forcément à l'utilisateur de choisir exactement l'emplacement de ses données : au mieux il a le choix entre différents centres de données certifiés ou agréés mais ces derniers ne se trouvent pas toujours dans la zone géographique voulue par l'utilisateur. Évidemment, ces limites sont cohérentes, les agréments et certifications n'existent que pour répondre aux exigences législatives des pays dans lesquels ils sont en place. Cependant, les utilisateurs souhaitant placer des données pour des critères de sécurité ou de performances, sans contraintes légales, sont contraints de faire confiance leur fournisseur.

De plus, sans mécanismes particuliers permettant d'assurer la localisation, et même si le fournisseur est de bonne foi, des erreurs de placement peuvent amener les données à être stockées dans une zone géographique non acceptable pour l'utilisateur. Dans le cas où le fournisseur ne serait pas de bonne foi, voire malveillant, il peut aussi les déplacer, dans une optique d'optimisation des coûts, sans en informer l'utilisateur.

C'est pourquoi il sera présenté des méthodes de localisation des données dont l'implémentation est à la charge du fournisseur dans ce chapitre et plus précisément des méthodes permettant d'assurer la bonne localisation des données. Seront présentées d'abord les méthodes nécessitant l'utilisation de matériel sécurisé puis celles nécessitant l'utilisation de logiciel sécurisé, avant de voir les limites de ce type de méthodes.

2.2 Utilisation de matériel sécurisé

Certaines méthodes utilisent du matériel sécurisé. Ce matériel repose le plus souvent sur un module de plateforme sécurisée (TPM ou Trusted Platform Module) qui permet de construire un environnement pour assurer la localisation.

2.2.1 Description d'un TPM

D'après les spécifications du Trusted Computing Group (TCG) [21], la description et le fonctionnement d'un TPM peut être expliqué de la manière suivante.

Qu'est-ce qu'un TPM ?

Un TPM est un composant matériel basé sur un processeur spécialisé, permettant la génération, le stockage et la manipulation de clés cryptographiques ainsi que la gestion de certificats. L'état du TPM doit être séparé du reste du système auquel il est connecté (système hôte), c'est-à-dire s'exécuter dans un environnement qui ne doit pas être accessible par le reste du système pendant l'exécution des routines du TPM.

La seule manière d'interagir entre le TPM et le système hôte est l'utilisation de l'interface définie par la spécification du TCG. Pour que ceci soit respecté il y a deux manières de mettre en place un TPM :

— Un système de type « System on a Chip » (SoC ou Système sur une puce) dans lequel tout le nécessaire pour le TPM repose dans un unique composant, c'est-à-dire que la RAM, la ROM, la mémoire Flash et le processeur nécessaire à l'exécution sécurisée des routines du TPM sont tous inclus dans la puce. Le composant échange avec le système hôte uniquement les commandes définies par l'interface et leurs résultats, et ce au travers d'un bus. La modification directe de l'état interne du TPM n'est donc pas possible.

— Une exécution en ressources partagées avec le système hôte. C'est-à-dire que le processeur et la mémoire du système hôte exécutent eux même les routines du TPM. Cependant, la mémoire dédiée au TPM n'est utilisable que quand le processeur se trouve dans un mode d'exécution particulier. Durant ce mode d'exécution, les zones mémoires réservées au TPM contiennent les points d'accès vers les routines et les zones inscriptibles du TPM. Grâce à cela, la modification directe de l'état interne du TPM

n'est pas possible, elle ne peut se faire qu'au travers des commandes définies par la spécification.

Un TPM est composé de différents modules spécialisés, comme montré sur la figure 2.1.

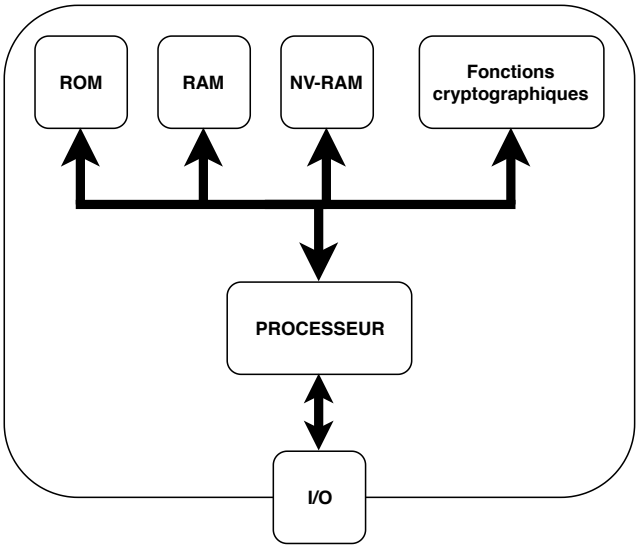


Figure 2.1 – Représentation de la structure d'un TPM

En partant de ce composant comme fondation, il est donc possible de construire des applications de confiance.

Différences entre la norme TPM 1.2 et 2.0

Deux normes de TPM cohabitent, la norme 1.2 dont la dernière spécification date de mars 2011 [22] et la norme 2.0 dont la dernière spécification date de septembre 2016 [21]. Même si les deux normes répondent aux mêmes cas d'utilisation et ont des fonctionnalités similaires, des différences existent. Une partie de ces différences se trouve au niveau de l'implémentation matérielle des TPM et n'est donc pas visible par l'utilisateur. Les autres différences concernent :

— La possibilité d'utiliser de nouveaux algorithmes de chiffrement et de hashage. La version 1.2 était limitée à RSA et SHA1. La version 2.0 permet l'utilisation de nouveaux algorithmes asymétriques, notamment d'algorithmes de cryptographie sur les courbes elliptiques ainsi que d'algorithmes de chiffrements symétriques comme AES.

— L'unification et l'amélioration de méthodes d'autorisation. Elles étaient différentes selon les cas d'utilisation en version 1.2. La version 2.0 fournit un cadre unifié. De plus, la version 2.0 permet d'utiliser des mots de passes et des Hash Message Authentication Code (HMAC) 1.2 comme nouvelles méthodes d'autorisation, en plus des politiques qui couvrent les méthodes proposées par la norme 1.2. Ces méthodes peuvent être combinées entre elles pour fournir une politique d'autorisation plus complexe.

— Support du TPM depuis le BIOS. En version 2.0, le support du TPM par le système d'exploitation n'est plus obligatoire pour être en mesure de l'utiliser.

— La mémoire RAM non volatile (NV-RAM) de la puce possède de nouvelles possibilités d'utilisation. En plus de stocker les clés et certificats, il est possible de s'en servir comme compteur ou tableaux de bits.

La version 2.0 permet donc d'être plus à jour en termes de sécurité, grâce à des algorithmes de chiffrement et de hashage plus récents et sans failles connues à ce jour et de faciliter l'utilisation du TPM, tout en conservant des cas d'utilisations similaires.

2.2.2 Méthodes garantissant une localisation exacte

Par l'utilisation d'un TPM en association avec une autorité de certification

Dans ce type de méthode [23], le principe est d'utiliser le TPM comme une « ancre de confiance pour valider la position géographique de machines virtuelles ». Une autorité tierce (TP ou Third Party), de confiance, est présente pour :

1. Certifier la bonne position géographique du TPM lors de la mise en place des machines physiques chez le fournisseur et vérifier par des audits réguliers que les TPM sont bien situés à l'endroit de départ et n'ont pas été déplacés.
2. Certifier la position géographique d'une machine virtuelle nouvellement crée.
3. Assister l'utilisateur durant la phase de vérification, en lui garantissant la validité des résultats retournés par le fournisseur.

L'architecture de la méthode est représentée sur la figure 2.2. La méthode est divisée en deux phases, durant ces phases toutes les communications entre

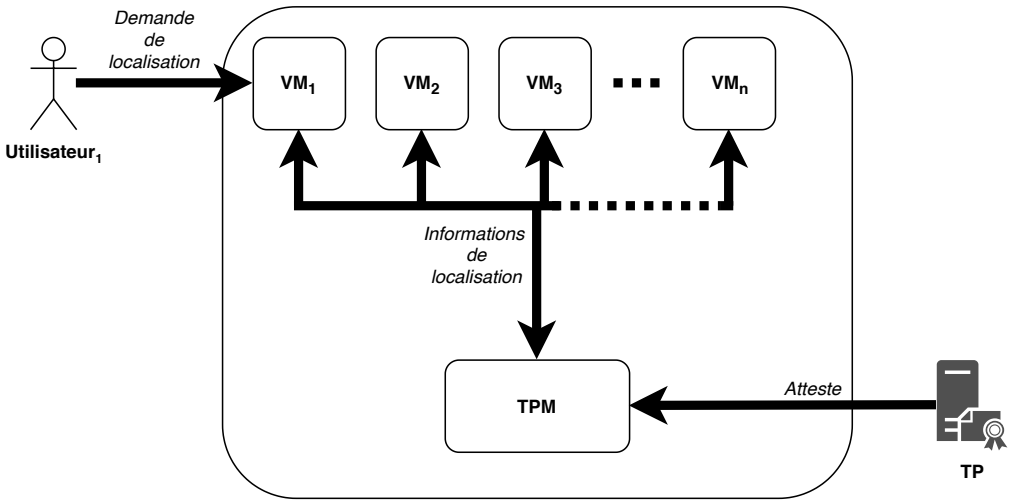


Figure 2.2 – Représentation de l'architecture

utilisateur, TP et fournisseur sont considérées comme chiffrées. Les phases sont les suivantes :

— Une phase d'initialisation durant laquelle le fournisseur déclare à la TP la position géographique de la nouvelle machine virtuelle en accord avec le TPM. Pour cela des échanges sécurisés ont lieu, permettant à la TP d'attester de la validité de la configuration matérielle de la machine hôte et au fournisseur de déclarer la position géographique et les éléments nécessaires à la vérification de la validité de la configuration matérielle.

— Une phase de vérification durant laquelle l'utilisateur interagit avec le fournisseur et la TP, afin d'obtenir la position géographique de sa machine virtuelle. Des échanges sécurisés entre ces trois entités ont lieu lorsque l'utilisateur génère une requête de localisation. D'abord, la TP identifie de qui provient la requête afin de retrouver les éléments nécessaires à la vérification. Une fois ces informations récupérées depuis le TPM chez le fournisseur, la TP atteste leur validité, et si tout est valide, la position enregistrée lors de l'initialisation est retournée à l'utilisateur. Ensuite, c'est à l'utilisateur de décider si la position géographique lui convient ou pas.

Par l'utilisation de récepteurs GPS

Cette méthode [24] repose sur l'utilisation d'un récepteur GPS « inviolable » au sein du centre de stockage. Le détail du récepteur GPS inviolable n'est pas

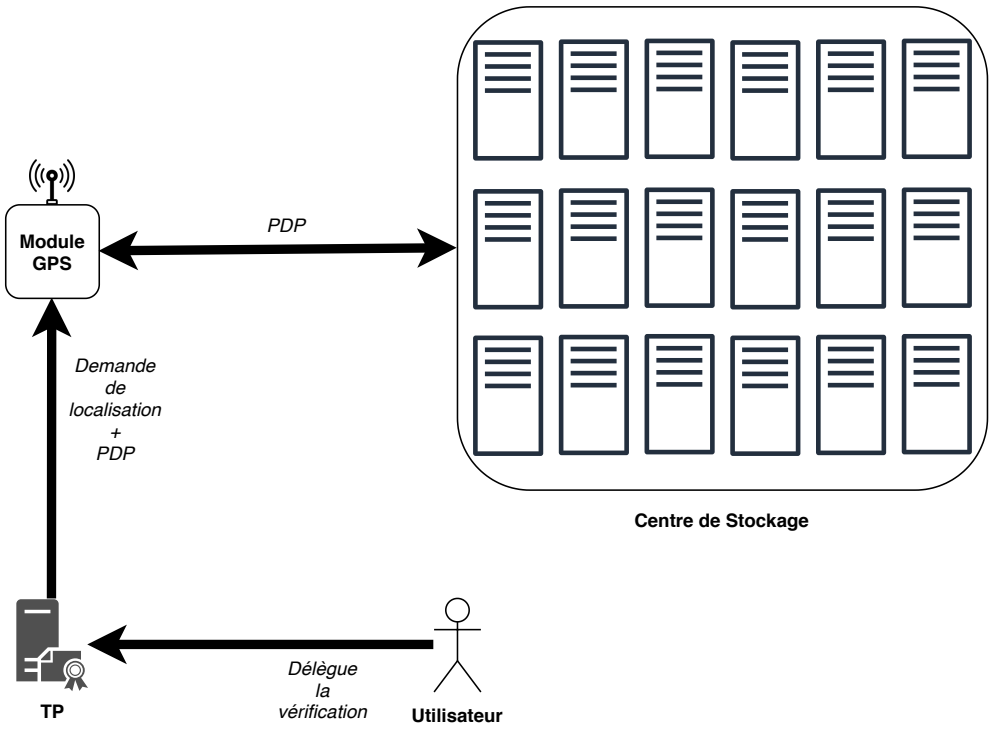


Figure 2.3 – Représentation de l'architecture

donné, mais on peut supposer que l'implémentation sécurisée est réalisée par un TPM. Le récepteur GPS est manipulé par une tierce personne (TP) et réalise la combinaison d'une preuve de possession de données (PDP) et d'un protocole liant le délai à la distance afin de s'assurer que les données se trouvent dans le centre de stockage.

Une PDP consiste à envoyer une série de défis/réponses au prétendu possesseur des données, ici le fournisseur, qui doit y répondre. Ces challenges consistent en des calculs cryptographiques sur les données, généralement un hash, produisant un « tag ». L'initiateur de la preuve connait la valeur du tag et demande au prétendu possesseur des données de réaliser le calcul afin de comparer le tag avec la valeur connue. Les PDP permettent d'obtenir une garantie probabiliste de la possession des données [25].

Le protocole liant le délai à la distance permet ici de savoir, en fonction du temps écoulé entre une requête et sa réponse, si le serveur qui répond à la requête est bien situé au sein du centre de stockage dans lequel se trouve le récepteur GPS ou bien se trouve ailleurs. De plus, il est à noter que le récepteur se trouvant dans le centre de données, il peut être directement connecté au réseau local dont la topologie, la vitesse et l'occupation des liens est connue. Les conditions sont donc optimales pour estimer le temps maximum entre

une requête et sa réponse. L'architecture de la solution est représentée sur la figure 2.3.

Une fois que le récepteur GPS est mis en place et que le temps maximum autorisé au sein du réseau est établi par la TP, la méthode est divisée en deux phases :

— Une phase d'initialisation, durant laquelle les fichiers sont séparés en blocs et le tag correspondant à chaque bloc est calculé par l'utilisateur. Le tout est envoyé au fournisseur par des moyens classiques, c'est-à-dire par le biais de l'interface mise à disposition par le fournisseur pour l'utilisateur.

— Une phase de vérification, initiée par l'utilisateur mais qui délègue sa réalisation à la TP. La TP envoie une requête de vérification de n blocs au récepteur GPS, qui choisit aléatoirement quels n blocs vérifier. En demandant chaque bloc et son tag au fournisseur le récepteur GPS mesure le temps avant réception du bloc et du tag. La liste des temps, les blocs et les tags correspondants ainsi que la position du récepteur GPS sont ensuite envoyés à la TP, de manière sécurisée. La TP vérifie la validité du temps maximum dans la liste par rapport à celui établi au départ. Après cela, les tags des blocs sont calculés et comparées à ceux reçus. Si ces deux données sont valides, la position renvoyée est validée, sinon elle ne l'est pas.

2.2.3 Méthodes garantissant le non-déplacement des données

Pour garantir le non-déplacement des données hors d'une ou plusieurs zones géographiques determinées [26], celles ci sont chiffrées et ne sont disponibles en clair que pour les centres de données situés dans les zones géographiques autorisées par l'utilisateur. Pour cela, chaque machine physique est équipée d'un TPM et d'un récepteur GPS, et un composant logiciel se charge d'enregistrer dans le TPM le lieu correspondant aux coordonnées géographiques de chaque machine, à intervalles réguliers. Ainsi, les informations de localisation sont stockées de manière fiable et sécurisée. Le TPM sert aussi, avec une TP, à attester de la validité logicielle des machines, c'est-à-dire si l'ensemble des logiciels présents sont considérés comme de confiance et ne vont pas contrecarrer le mécanisme de localisation.

La TP génère un couple de clé publique / clé privée et créé un hash pour chaque machine du fournisseur, en fonction de la position géographique initiale et de la validité logicielle de la machine. L'architecture de la solution est

résumée par la figure 2.4.

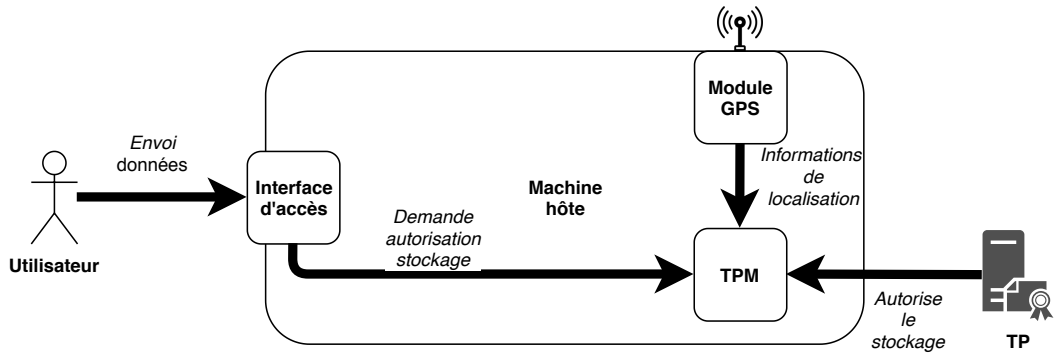


Figure 2.4 – Représentation de l'architecture

Le protocole se déroule ensuite de la manière suivante :

1. L'utilisateur chiffre avec un protocole symétrique et une clé K ses données. La clé K est ensuite chiffrée par la clé publique de la TP. L'utilisateur envoie ses données chiffrées, la clé K chiffrée et les lieux de stockage voulus au fournisseur.

2. À la réception de la requête de stockage, le fournisseur redirige les données chiffrées vers les hôtes dans les zones géographiques données par l'utilisateur. Ensuite le hash de chaque machine ayant reçue les données chiffrées est envoyée à la TP, ainsi que la clé K chiffrée.

3. La TP vérifie la validité logicielle et que la position est correcte, en fonction du hash précédemment stocké pour chaque machine. Le TP déchiffre K et l'envoie aux machines ayant réussi la vérification.

4. Les hôtes ayant reçu la clé K déchiffrée s'en serve pour déchiffrer les données et les stocker en clair.

Il existe aussi des méthodes similaires [27], sans TP, où les échanges sont réalisés uniquement entre l'utilisateur et le fournisseur, mais n'assurant la bonne localisation que si le fournisseur n'est pas malveillant et implémente correctement la méthode.

2.3 Politiques au sein de la pile logicielle du fournisseur

2.3.1 Méthodes estimant une localisation

Une façon d'estimer la localisation des données au niveau logiciel est d'encapsuler les données dans des scripts [28]. Les données ne sont ainsi plus des éléments passifs du système, mais deviennent des éléments actifs et exécutables. Pour permet « l'exécution des données », un système de fichier spécifique est utilisé. L'accès aux données n'est donc possible qu'en communicant avec la couche extérieure, les scripts, et défini selon des politiques d'accès. Du point de vue de l'utilisateur, l'utilisation est transparente, l'encapsulation étant réalisée par le fournisseur. Les différents échanges entre l'utilisateur et les données se font de manière sécurisée. Pour que cette méthode puisse fonctionner correctement, le fournisseur est considéré comme de confiance et n'altère ni l'éxecution des scripts ni le système en général. À l'initialisation, l'utilisateur et les données sont représentés par un ensemble de paramètres qui définira la manière dont ils communiqueront ensuite. L'architecture de cette méthode est représentée sur la figure 2.5.

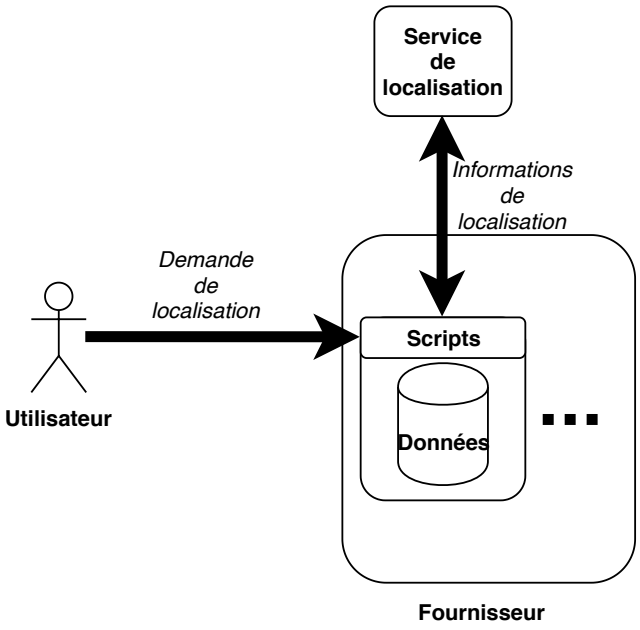


Figure 2.5 – Représentation de l'architecture

Les données acceptent différentes requêtes, dont une concernant leur locali-

sation physique, qui permet à l'utilisateur de demander aux données de donner leur localisation géographique. Cette opération se déroule de la manière suivante :

1. L'utilisateur envoie la requête de localisation au fournisseur.
2. À la réception de la requête, le fournisseur recherche les données concernées et leur transmet la requête.
3. Les données ayant reçu la requête vérifient leur provenance grâce aux paramètres échangés lors de l'initialisation et si tout est valide, elles peuvent exécuter l'opération de vérification de la localisation, sinon elles arrêtent leur exécution ici.
4. Pour connaître leur localisation les données envoient une requête à un service permettant d'obtenir la localisation en fonction des informations sur la topologie contenue dans la requête.
5. Le résultat est renvoyé au fournisseur, qui le renvoie ensuite à l'utilisateur.
6. De la même manière que les données ont vérifié la validité de la requête, l'utilisateur vérifie la validité de la réponse. Si tout est valide, la localisation est confirmée, sinon elle ne l'est pas.

De plus, si les données sont transférées vers une autre localisation, elles peuvent exécuter d'elles-mêmes l'opération de vérification de la localisation, afin d'informer automatiquement l'utilisateur de la nouvelle localisation. En supposant le service de localisation fiable, la localisation peut ainsi être assurée à l'utilisateur.

2.3.2 Méthodes garantissant le non-déplacement des données

Le but de ses méthodes est d'assurer le non déplacement des données dans une zone géographique non autorisée par l'utilisateur. Elles nécessitent la mise en place de logiciels spécifiques chez le fournisseur. Elles fonctionnent grâce à des politiques sur les actions autorisées et non autorisées au niveau des données, par exemple le maintien dans une zone géographique spécifique]. Les politiques sont fournies par l'utilisateur et évaluées au sein de la pile logicielle du fournisseur. L'implémentation peut se faire au niveau du système de fichier [29] ou à plus haut niveau [30, 31].

Ces politiques sont évaluées avant de réaliser chaque opération sur les fichiers et l'opération n'est exécutée que si l'état du fichier après exécution de

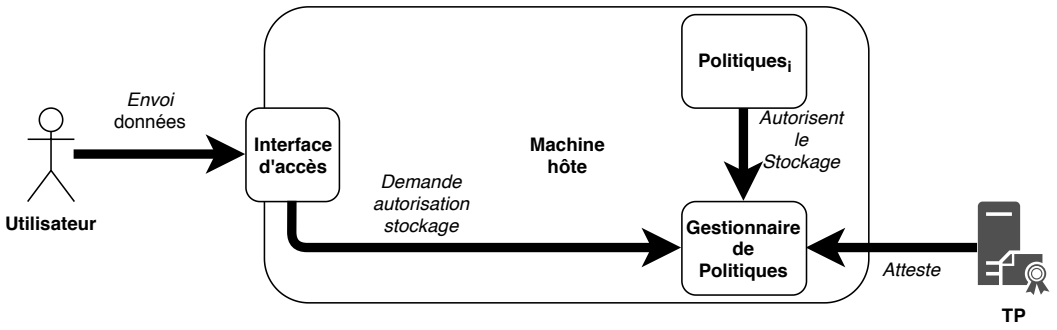


Figure 2.6 – Représentation de l'architecture

l'opération respecte les conditions définies par la politique. L'architecture de ces solutions est représentée sur la figure 2.6.

La méthode fonctionne en deux étapes :

— Une étape d'initialisation, durant laquelle l'utilisateur définit les politiques qu'il souhaite associer à ses fichiers, et notamment la position géographique souhaité. Les politiques sont ainsi envoyées au fournisseur, selon une interface spécifique pour ne pas les confondre avec les données de l'utilisateur. La localisation des centres de stockage est donc connue par l'utilisateur, afin qu'il puisse choisir au moins un lieu où se trouve un centre de stockage dans ses politiques de localisation. L'utilisateur peut ensuite commencer à déposer ses fichiers.

— Une étape de vérification, s'exécute avant chaque opération sur les fichiers, quel que soit le type d'opération. Le déroulement de cette étape est le suivant :

1. Les politiques associées à l'utilisateur sont récupérées.
2. L'état du fichier après l'opération à réaliser (état final) est calculé, mais l'opération n'est pas réellement exécutée.
3. L'état du fichier après l'opération à réaliser est comparé aux politiques.
4. Si l'état final respecte les politiques, l'opération peut être réalisée, sinon elle ne l'est pas.

Selon les méthodes, les journaux des opérations exécutées et non-exécutées sont conservés. Une TP peut aussi intervenir afin de réaliser des audits réguliers pour veiller au bon fonctionnement du système et au respect des politiques par le fournisseur.

2.4 Limites de ces méthodes

Ces méthodes fournissent la possibilité d'offrir à l'utilisateur une garantie sur la position de ses données, sous la condition qu'elles soient correctement implémentées. Cependant même dans ce cas elles présentent des limites, qui seront abordés dans la suite de cette section.

2.4.1 Coût élevé

La première limite de ces méthodes est le coût de mise en œuvre. En effet, dans le cas le moins couteux, mais aussi le moins fiable, il faut « imposer » au fournisseur d'utiliser un logiciel mettant en œuvre les mécanismes permettant l'assurance. En supposant le fournisseur de bonne foi, il doit accepter de déployer ce logiciel, ce qui peut être contraignant pour lui sur plusieurs niveaux :

— Au niveau de l'administration, car le logiciel doit être déployé et intégré au sein de la couche logicielle existante et maintenu au fil des évolutions.

— Au niveau des manipulations des données, car tout doit être tracé au niveau de la position physique.

— Au niveau de la performance car chaque opération étant tracée, il faut réussir à assurer la même QoS en utilisant plus de ressources.

Dans le cas le plus coûteux, le fournisseur doit installer du matériel sécurisé sur l'ensemble de ses machines. En effet, même si les TPM sont des composants courant ils ne sont pas tous de la norme 2.0 qui permet d'avoir une sécuritée renforcée. Il faudrait donc installer des TPM issus de la norme 2.0 sur l'ensemble des machines. De plus il doit aussi collaborer avec une TP afin de prouver sa bonne foi.

Toutes ces contraintes sur le fournisseur se feront ressentir financièrement sur l'utilisateur. Selon le type d'utilisateur, ce n'est pas toujours acceptable.

2.4.2 Failles

La seconde limite de ces méthodes sont les failles qui peuvent exister à différents niveaux. Il est à noter que ces failles peuvent être impossible à exploiter si un audit régulier par une TP est mis en place, à moins que le fournisseur soit malicieux et dispose de ressources quasi-infinies lui permettant de masquer les failles, c'est-à-dire déplacer les données dans à la bonne position géographique,

modifier les journaux et avoir le matériel correctement configuré, avant chaque audit. Cette hypothèse peut donc être facilement exclue.

Failles au niveau logiciel

Toujours en considérant que le logiciel est correctement implémenté, s'il n'est pas utilisé en coopération avec un TPM et une TP, rien ne garantit à l'utilisateur que le logiciel qui est en train d'être exécuté chez le fournisseur est bien celui permettant d'assurer la localisation. En effet l'utilisateur n'a aucun accès physique aux hôtes et aucune autre autorité de confiance ne lui garantie la bonne exécution du logiciel de localisation. Cette absence du logicel de localisation peut être une erreur du fournisseur ou bien l'acte d'un fournisseur malveillant.

Failles au niveau GPS

Certaines méthodes utilisent des récepteurs GPS au sein des centres de données. Cependant cela pose deux problèmes majeurs :

— La localisation exacte d'un centre de données n'est pas toujours souhaible pour un fournisseur. En effet, les fournisseurs souhaitent généralement la garder secrète pour des raisons de sécurité, notamment afin d'éviter les actes de malveillance. Les fournisseurs peuvent ainsi se montrer réticent à l'utilisation d'un tel système fournissant une localisation aussi précise.

— Les signaux GPS authentiques, provenant des satellites, peuvent être usurpés par des faux signaux générés afin de faire croire à une autre position géographique [32]. Ce sont des mécanismes qui peuvent être mis en place par un fournisseur malveillant, afin de faire croire qu'une machine contenant des données devant se trouver au sein d'une zone géographique s'y trouvent encore alors qu'elles ont été déplacées.

Failles au niveau TPM

Les TPM, censés assurer la fiabilité du système grâce à leurs capacités cryptographiques, ne sont pas exempts de failles. Des attaques permettant de récupérer les secrets stockés au sein du module existent [33,34], mais leur mise en œuvre est difficile et coûteuse pour un fournisseur.

Cependant une autre faille plus facile à exploiter est connue [35]. Il s'agit d'une faille de type ROCA (Return of Coppersmith's attack) qui se situe au niveau du code permettant la génération des clés au sein du TPM. La clé privée

peut être déduite de la clé publique, permettant ainsi à l'attaquant d'utiliser les clés privées normalement non accessibles.

Un fournisseur malicieux pourrait exploiter ce type de failles afin de masquer la localisation réelle.

Chapitre 3

Estimation de la localisation dans le Cloud à l'aide de points de repère

Sommaire

3.1 Introduction

Des méthodes permettant de garantir la localisation existent, comme présentées dans le chapitre 2 et même si elles sont sensibles à certaines failles, celles-ci peuvent être évitées en ayant recours à une entité tierce, reconnue comme de confiance par le fournisseur et l'utilisateur.

Cependant, ces méthodes présentent un coût élevé pour le fournisseur, et encore plus si une entité tierce intervient dans la mise en place. Ce coût étant retranscrit financièrement au niveau de l'utilisateur, il n'est ainsi pas toujours possible d'y recourir. En effet, tous les utilisateurs n'ont pas le même profil et la même capacité financière. De plus, tous les fournisseurs ne souhaitent pas employer de telles méthode.

C'est pourquoi, d'autres méthodes fondées sur la mise en place de points de repère par l'utilisateur et leur collaboration existent afin non pas d'assurer mais d'estimer la position géographique des données. C'est un bon compromis entre d'une part les méthodes « fiables », mais coûteuses, assurant une localisation précise des données, et d'autre part l'absence de processus de contrôle nécessitant d'accorder sa confiance au fournisseur (en espérant qu'il ne soit ni malveillant ni sujet aux erreurs de placement de données). En effet, même si la position géographique n'est qu'estimée, le coût de mise en œuvre est faible et ne nécessite pas un déploiement constant des points de repère pour l'utilisateur.

Un point de repère est simplement une machine connectée à Internet et dont la position est connue. Cette machine est généralement contrôlée par l'utilisateur et à un accès au service Cloud dont l'utilisateur souhaite connaître la position géographique, c'est-à-dire que le point de repère est en mesure d'utiliser l'interface fournie par le fournisseur du service afin d'interroger les données.

Dans ce chapitre, nous présentons des méthodes de localisation des données dont l'implémentation est à la charge de l'utilisateur, et plus précisément des méthodes permettant d'estimer la bonne localisation des données. Les différents critères permettant de classifier ces méthodes puis les résultats expérimentaux revendiqués par les auteurs seront présentés, avant de voir les limites de ce type de méthodes. L'ensemble de ce chapitre reprend des travaux précédemment publiés [36, 37].

3.2 Fonctionnement général des méthodes

L'ensemble de ces méthodes fonctionne généralement de la même manière. L'idée principale est qu'en apprenant avec précision les performances du réseau

autour de la zone présumée dans laquelle le fournisseur stocke ses données, dans un premier temps sans interagir avec le fournisseur, il est ensuite possible de retrouver plus précisément cette zone. En effet, les performances du réseau doivent être proches ou identiques à celles observées lors de l'apprentissage. La zone estimée sera potentiellement plus précise si elle est entourée de points de repères que si elle se trouve au sein d'une autre zone géographique, non couverte pas les points de repères. Il s'agit donc de considérer les deux étapes correspondantes aux méthodes, une étape d'apprentissage des performances du réseau à l'aide de métriques données et une étape de vérification utilisant les résultats de l'étape d'apprentissage afin d'inférer la position géographique du fournisseur. L'architecture et le fonctionnement général de ces méthodes est représentée sur la figure 3.1.

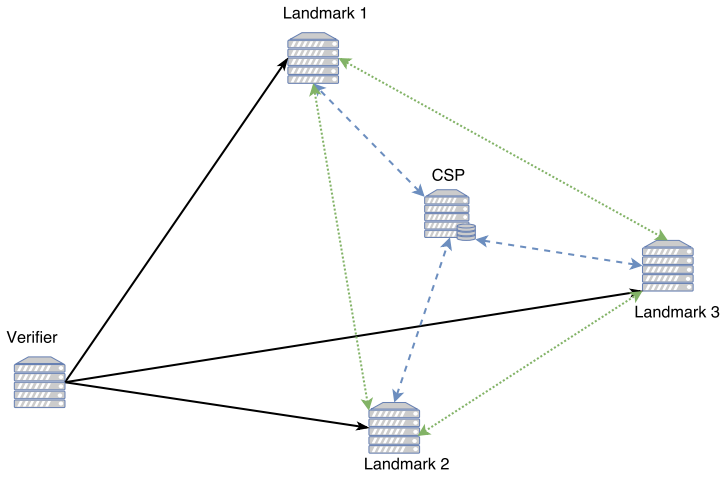


Figure 3.1 – Représentation de l'architecture

3.2.1 Étape d'apprentissage

La première étape est celle d'apprentissage, durant laquelle les points de repères interagissent entre eux afin de collecter des mesures du réseau comme les Round Trip Times (RTTs) ou le nombre de sauts (HC) entre deux points de repère. Dans la plupart des cas, seuls les RTTs sont considérés, ainsi, « mesures du réseau » se référera aux RTTs uniquement, sauf mention contraire. Ces mesures du réseau, entre chaque couple de points de repère, sont ensuite utilisées afin de déterminer les paramètres d'un modèle d'apprentissage automatique, déterminé à l'avance et servant à l'estimation de la zone géographique. L'étape d'apprentissage correspond donc à la phase de collecte des données et à leur utilisation pour alimenter un modèle d'apprentissage automatique.

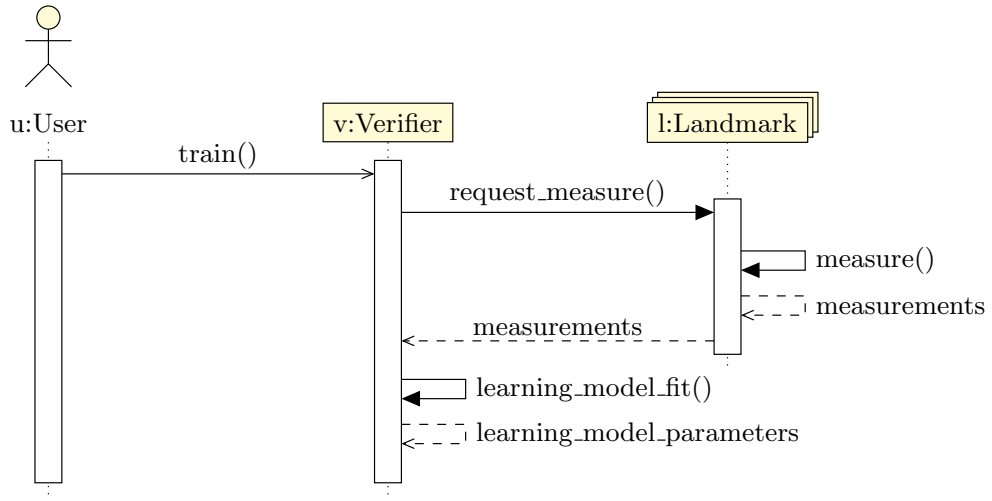


Figure 3.2 – Étape d'apprentissage (centralisée)

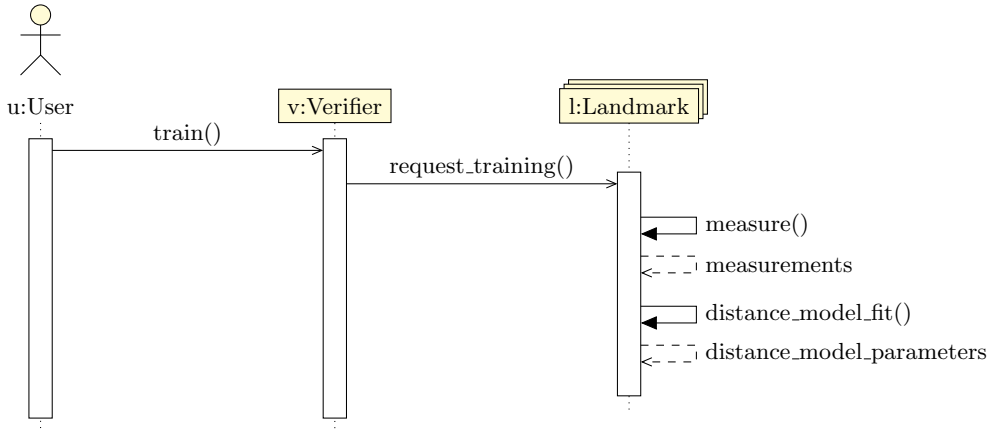


Figure 3.3 – Étape d'apprentissage (décentralisée)

3.2.2 Étape de vérification

La deuxième étape est celle de vérification, lancée à la demande de l'utilisateur qui souhaite estimer la position de ces données. Les points de repères sont notifiés de la vérification et chacun interagit avec le fournisseur stockant les données, en réalisant soit des simples « ping » soit des accès aux données, afin de récolter des mesures du réseau concernant le fournisseur. En utilisant ces mesures nouvellement recueillies et le modèle d'apprentissage automatique précédemment établi et dont les paramètres ont été déterminés lors de l'étape d'apprentissage, une zone d'estimation de la localisation du fournisseur peut être calculée.

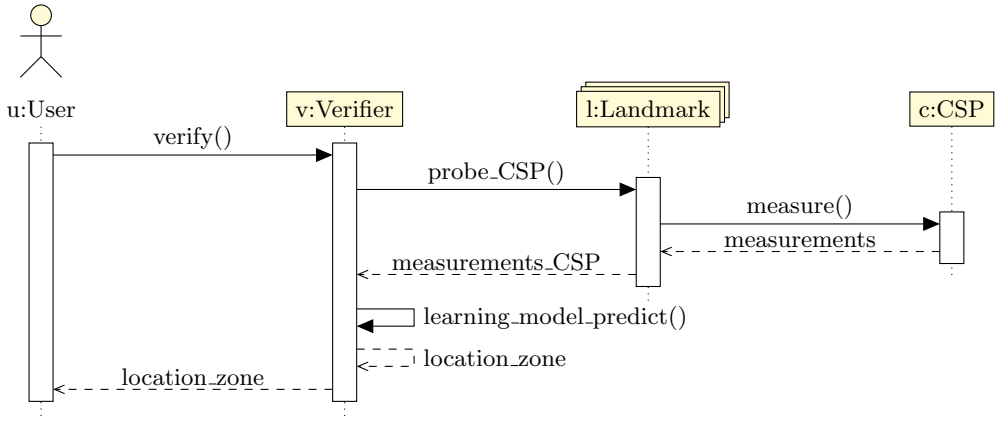


Figure 3.4 – Étape de vérification (centralisée)

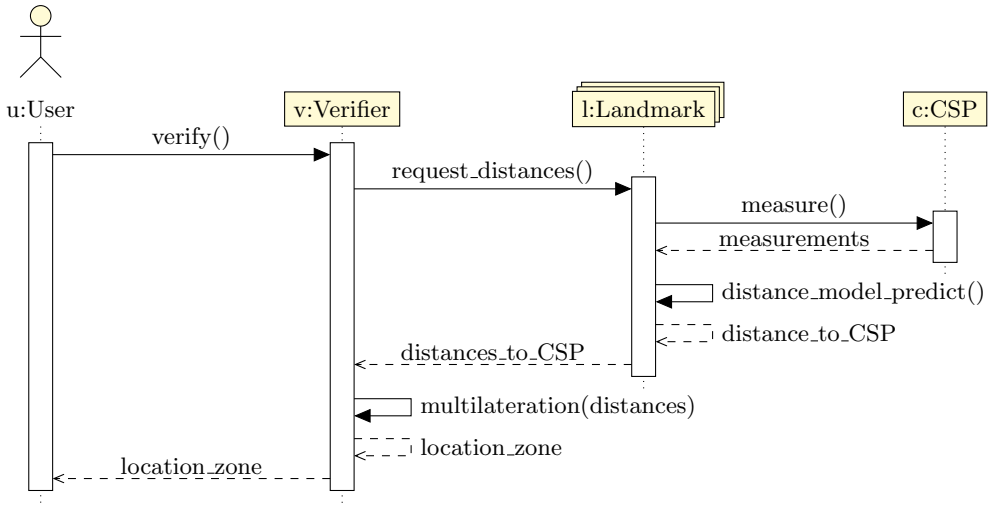


Figure 3.5 – Étape de vérification (décentralisée)

3.3 Classification des méthodes

Les différentes méthodes existantes différent selon de multiples critères de conception. Ces critères sont présentés dans la suite et résumés dans le tableau 3.1. Ils peuvent être regroupés sous cinq catégories : Corrélation entre l'adresse du point d'accès et la position du serveur (Corrélation @PA/Pos.), points de repères, recueil de mesures, apprentissage automatique et utilisation de protocoles de PDP.

3.3.1 Corrélation entre l'adresse du point d'accès et la position du serveur

Les utilisateurs utilisent un point d'accès particulier pour manipuler leurs données stockées dans le Cloud. Ce point d'accès peut être sous la forme d'une adresse IP ou un nom de domaine et il est fourni à l'utilisateur lors de l'établissement du contrat du service. Les méthodes présentées ont deux points de vue sur la corrélation @PA/Pos. :

— Existence d'une telle corrélation [38–40] : L'adresse du point d'accès utilisée par l'utilisateur est celle des serveurs stockant les données ou d'un proxy mais situé dans le même centre de données que les serveurs de stockage. Ainsi, en termes de vérification de la localisation, sonder l'adresse du point d'accès revient à sonder les serveurs stockant les données.

— Non-existence d'une telle corrélation [41–45] : L'adresse du point d'accès fournie n'est pas celle des serveurs stockant les données mais d'un proxy qui peut être situé loin des données. Ainsi, en termes de vérification de la localisation, l'adresse du point d'accès n'est pas utile, car elle ne servira qu'à estimer la localisation du proxy et non des données.

Il est important de noter que les fournisseurs commerciaux, dont les infrastructures reposent essentiellement sur de la virtualisation, sont plus susceptibles de se situer dans la seconde situation, afin de rendre opaque l'architecture interne de leur infrastructure.

3.3.2 Points de repères

Cette catégorie peut être divisée en trois critères de conception :

Types de points de repères

Un point de repère peut être de deux types différents selon la manière dont il est employé au sein de l'ensemble du processus de vérification.

— Les points de repères actifs [38–44] sont a l'initiative des requêtes dans le but de collecter des mesures du réseau. Ils peuvent aussi mettre en place un processus d'apprentissage et déduire des distances à partir des RTTs. Ce sont des machines dont la puissance de calcul peut être directement utilisée.

— Les points de repères passifs [45] répondent uniquement aux requêtes qui leur sont adressées et servent à collecter les RTTs, mais leur puissance de calcul n'est pas directement exploitable.

Échelle de distribution des points de repère

L'échelle de distribution des points de repère fait référence à la zone dans laquelle sont situés les points de repère. Une telle zone peut être :

— La planète entière pour une échelle mondiale comme utilisés par [39, 40, 45].

— Un ou plusieurs continents ou un très grand pays pour une échelle continentale. Par exemple [38, 41–43] positionnent les points de repère aux États-Unis et [44] positionne les points de repère aux États-Unis et en Europe.

— Un petit pays ou une partie d'un pays, comme un état, une région ou un conté, pour une échelle locale.

Sélection de points de repère

Un ensemble de points de repère est défini à la mise en place de la solution et le rôle individuel des points de repère est de contribuer à déterminer la zone géographique dans laquelle se situe le fournisseur. Néanmoins, tous les points de repère de l'ensemble initial ne peuvent pas contribuer de la même manière et apporter plus de précision. Par exemple, un point de repère trop loin de tous les autres ou du fournisseur va au mieux n'avoir aucun intérêt dans l'estimation de la localisation et au pire impacter négativement la zone géographique déterminée. La sélection des points de repère cherche donc à trouver un compromis entre la couverture géographique et la précision de l'estimation. En effet, un nombre important de points de repère et répartis mondialement permet de localiser à la même échelle, mais si tous les points de repères participent à la

vérification, l'estimation risque de ne pas être précise. Sans prendre en compte les cas où des points de repère sont écartés pour des raisons techniques, il est possible de distinguer deux paradigmes de sélection :

— Pré-apprentissage [45], pour sélectionner les points de repère les plus proches de la localisation présentie, afin de minimiser les coûts, et de ne pas utiliser des points de repère qui n'apporteront rien ou très peu au modèle d'estimation. Cette sélection se justifie par le nombre élevé de points de repère dans l'ensemble initial.

— Pré-vérification [40, 41], aussi pour sélectionner les points de repère les plus proches de la localisation présentie. Cette sélection est utile lorsque le nombre de points de repère est limité dans l'ensemble initial. Leur nombre n'est donc pas un problème pour l'apprentissage, mais la sélection pré-vérification est utile pour améliorer la précision de la vérification en excluant les points de repère les plus lointains, qui n'auront pas d'influence positive sur le résultat.

Le reste des solutions [38,39,42–44] ne sélectionnent pas les points de repère au sein de l'ensemble initialement choisi, à part, des points de repère qui ne fonctionnent pas correctement. Cette sélection n'est pas comptabilisée, car elle ne fait pas partie de « l'esprit » de la solution mise en œuvre.

3.3.3 Collecte de mesures

Il y a deux aspects à prendre en compte lors de la conception d'une solution en termes de collecte de mesures.

Métriques

Le premier aspect à prendre à compte est les métriques à utiliser. La plus utilisée est le RTT, que ce soit les valeurs brutes, les moyennes sur un nombre de mesure prédéfini, la médiane, le mode ou bien l'écart type du RTT. Toutes ces mesures sont confondues sous le terme de RTT et toutes les solutions l'utilisent. Le nombre de saut (HC) et la bande passante (BW) entre un point de repère et la cible de sa sonde sont parfois collectées afin de renforcer la précision du modèle [38].

Protocole de sonde

Le deuxième aspect est la façon de collecter les métriques, c'est-à-dire le protocole utilisé lors de la sonde par les points de repère. Il y a principalement

deux manières de faire, en réalisant des simples « ping » ou « traceroute » qui utilisent tous les deux le protocole ICMP [38–41] ou bien en utilisant l'interface du fournisseur, fondée sur le protocole HTTP(S), afin d'accéder réellement aux fichiers [41–45].

Utiliser le protocole ICMP résulte en un RTT plus précis comparé à l'utilisation du protcole HTTP(S). En effet, en se reposant sur HTTP(S), la sonde réalise des accès aux fichiers et est soumise à un double « overhead ». Le premier étant celui du protocole HTTP(S) par rapport à ICMP, ce premier faisant partie de la couche application du modèle TCP/IP, une couche haute, alors qu'ICMP fait partie de la couche réseau, une couche basse, du même modèle. L'overhead lié à HTTP(S) est donc plus important que celui lié à ICMP. Le second overhead existant est celui existant lors de l'accès réel au fichier, cet accès prend plus de temps que de simplement interroger le serveur et créé plus de variation dans les RTTs mesurés, notamment à cause des mécanismes de mise en cache des données. Le protocole ICMP est utilisé dans le cas où une corrélation @PA/Pos. est considérée comme existante.

D'autre part, utiliser le protocole HTTP(S) garantie que la cible de la sonde possède bien les données que l'utilisateur essaye de localiser, c'est-à-dire que dans le cas où l'interface est un proxy, le proxy doit transmettre la requête au serveur contenant les données. Ce-faisant le serveur contenant les données est bien celui sondé, comparé à l'utilisation d'ICMP où le sondé peut être le proxy. Cette manière de sonder est utilisée dans le cas où une corrélation @PA/Pos. est considérée comme non-existante.

3.3.4 Apprentissage automatique

Coordination lors de l'apprentissage

Il y a deux catégories d'approches en ce qui concerne la coordination des points de repères lors de l'apprentissage :

— Les approches centralisées [38, 39, 42, 45] au sein desquelles les points de repère ne servent qu'à collecter les RTTs. Même si les points de repères utilisés sont « actifs », le modèle d'apprentissage automatique est centralisé au niveau du vérifieur, qui peut être l'utilisateur ou un hôte délégué à cette fonction. Il calcule donc les paramètres en fonction des valeurs des RTTs lors de l'apprentissage (figure 3.2) et applique lui-même le modèle en fonction des valeurs des RTTs reçues lors de la vérification (figure 3.4).

— Les approches décentralisées [40,41,43,44] au sein desquelles les points de repère sont forcément actifs et en plus de collecter les RTTs, établissent

chacun les paramètres d'un modèle d'apprentissage automatique lors de l'apprentissage (figure 3.3) et appliquent ce modèle pour estimer une distance lors de la vérification (figure 3.5).

Estimation de la distance

Lorsque les mesures du réseau sont réalisées lors de la vérification, les méthodes étudiées les utilisent pour inférer la distance entre le point de repère ayant réalisé la mesure et le fournisseur. Deux techniques ont été identifiées :

— Construire une fonction f calculant la distance d en fonction de mesures M selon différentes métriques, telle que $d = f(M)$. Plusieurs façons de choisir la fonction f existent :

  — Par régression linéaire [42, 44].

  — Par régression polynomiale [45].

  — En utilisant le ratio délai / distance [41], qui correspond à la moyenne des RTT mesurés sur la distance entre l'ensemble des points de repères.

  — En utilisant une bestline [40,43], c'est-à-dire la fonction linéaire avec les plus grandes pente et ordonnée à l'origine, de telle sorte que sur un graphe représentant les distances en abscisse et les RTTs en ordonnées, toutes les valeurs mesurées sont au-dessus de la ligne représentant la fonction [46].

  Il est à noter que dans le cas d'approches décentralisées, le choix de la fonction est réalisé à l'avance et chaque point de repère utilise la même fonction, seul les valeurs des paramètres changent. C'est-à-dire qu'ils peuvent utiliser des fonctions linéaires pour tous, ou bien des bestlines pour tous mais pas de combinaisons.

— Il est aussi possible d'utiliser des systèmes de coordonnées virtuels, associant chaque point de repère à un point dans un espace géométrique. La distance entre les points de l'espace géométrique représente leur distance en fonction du RTT mesuré. Cependant, les valeurs ne représentent pas les distances réelles mais correspondent directement au RTT mesuré entre deux points de repères. Le fournisseur est ensuite associé à un point dans l'espace géométrique. Par conséquent, la position ne peut être inférée que par classification. Différents systèmes existent comme Vivaldi [47], Pharos [48] et Phoenix [49]. La méthode [39] utilise le système

de coordonnées virtuelles Phoenix.

La méthode [38] n'utilise pas d'estimation de la distance et infère directement la position en fonction du RTT.

Inférence de la localisation

Le vérifieur réalise toujours la dernière étape, qui est l'inférence d'une zone géographique dans laquelle le fournisseur stocke ses données, en fonction des paramètres calculés lors de l'apprentissage et des mesures réalisées lors de la sonde du fournisseur. Deux méthodes existent :

— Souvent utilisées dans les approches centralisées, les techniques de classification permettent d'inférer la localisation, en établissant d'abord une liste de lieux connus et susceptibles d'abriter les données. Les mesures collectées lors de l'apprentissage servent à entraîner le classificateur et celles collectées lors de la vérification lui permettent de décider d'une localisation parmi celles possibles. Différents algorithmes sont utilisés, tels que la classification naïve bayésienne [38], la classification « Instance-Based » [39] et le regroupement hiérarchique [42].

— Dans les approches décentralisées, la multilatération est utilisée. Avant cette étape, les points de repère ont chacun estimé leur distance par rapport au fournisseur et l'ont envoyé au vérificateur. De plus les positions de chaque point de repère sont connues, par définition des points de repère. En supposant n points de repère, chacun peut être associé à sa paire de coordonnées (x_i, y_i) déterminant sa position et une distance d_i, celle qui a été calculée et envoyée au vérificateur. Il est possible de créer n cercles C_i, de centre (x_i, y_i) de rayon d_i. La multilatération est ainsi la fonction de M prenant en entrée les n cercles et retournant un polygone P, résultat de l'intersection des n cercles, telle que $P = M(C_0, C_1, ..., C_{n-1})$. Ensuite l'interprétation du polygone P peut être un pays, une ville, la réduction aux coordonnées d'un point particulier du polygone, etc. Un example de multilatération parfaite, résultant en un point est illustrée par la figure 3.6. La figure 3.7 illustre une multilatération surestimant les distances, résultant en une figure géométrique.

Granularité de la localisation

À la fin du processus de vérification, un résultat est donné à l'utilisateur. Une granularité est considérée fine quand le résultat est sous forme de coordonnées GPS [41,45], d'une ville [42] ou d'un conté [38]. À l'opposé une granularité

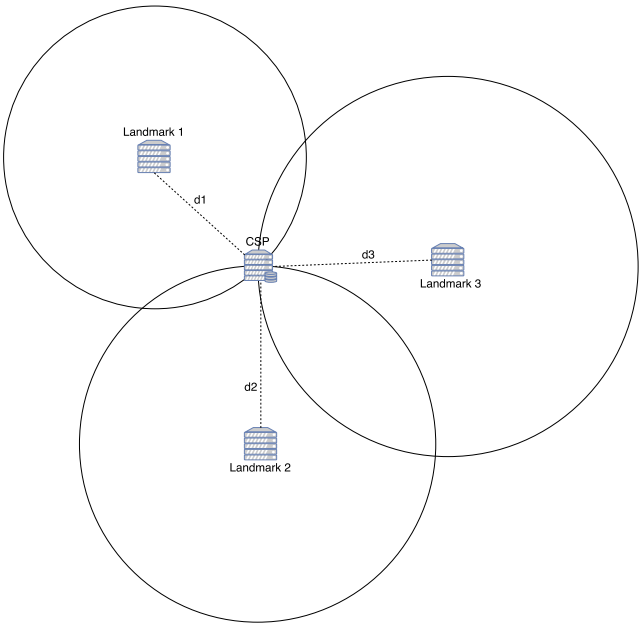


Figure 3.6 – Exemple de multilatération parfaite

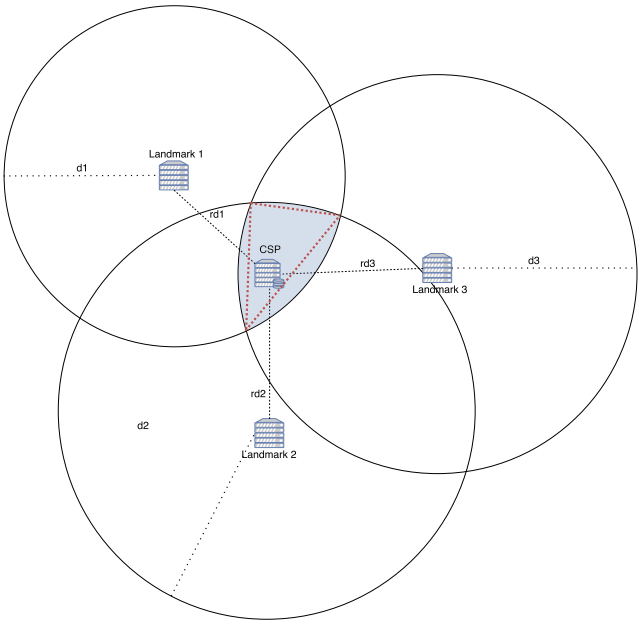


Figure 3.7 – Exemple de multilatération où les distances sont surestimées

est dite grossière lorsque le résultat est un pays [39] ou un continent.

Certaines méthodes [40, 43, 44] résultent en une granularité variable qui dépend essentiellement de la variance des mesures récoltées. Il est alors possible d'obtenir une granularité variant de très fine à très grossière.

La granularité est une propriété quantitative du processus de vérification, et tous les utilisateurs n'ont pas le même besoin en granularité. Certains ont besoin de stocker les données au sein d'un pays, et n'ont pas envie de supporter des coûts élevés pour recevoir un résultat de granularité plus fine.

Cependant, comme les processus de vérification fondent leur mise en œuvre sur de l'apprentissage, les estimations ont une erreur associée. Par conséquent, la probabilité que le fournisseur stocke les données dans la zone inférée est d'autant plus élevé que la zone est large. En partant de ce constat, une contrainte de granularité de localisation trop fine peut compromettre le processus de vérification en fournissant des faux négatifs. C'est-à-dire que l'interprétation du résultat serait « Le fournisseur ne stocke pas ses données dans une zone autorisée par l'utilisateur », alors que l'interprétation contraire aurait été possible si la zone était plus large.

3.3.5 Utilisation de protocoles de PDP

Des protocoles de PDP peuvent être mis en place dans les cas où le protocole de sonde est HTTP(S) afin de s'assurer que les données retournées sont bien les données demandées. Il est possible de distinguer deux façons de mettre en place ces protocoles en fonction de l'endroit où sont calculées les preuves :

— Si les preuves sont calculées côté utilisateur / vérifieur elles sont réalisées à l'aide d'un Message Authentication Code (MAC). Les fichiers sont découpés en blocs B_i et un tag t_i, identifiant le bloc i, est crée en calculant le hash du bloc avec une fonction H telle que $t_i = H(B_i)$. Les blocs sont stockés chez le fournisseur et les tags peuvent être soit conservés par l'utilisateur soit stockés eux aussi. Pour vérifier la possession des données, l'utilisateur envoie une liste de c numéros de blocs, sélectionnés aléatoirement, au fournisseur. Le fournisseur doit ensuite renvoyer les c blocs correspondants. L'utilisateur récupère les tags correspondants aux c blocs, calcule les tags des blocs reçus et les compare au tag précédemment calculés afin de conclure de la bonne possession des données ou non. Bien que ce protocole soit consommateur de bande passante, il ne nécessite pas d'action particulière du serveur et peut être facilement mis en place, comme dans les solutions [43, 44].

— Si les preuves sont calculées côté fournisseur / prouveur il est possible

d'utiliser une PDP cryptographique [50]. C'est-à-dire que l'utilisateur commence par créer un couple de clé privé / clé publique. Les fichiers sont ensuite découpés en blocs B_i et un tag t_i, identifiant le bloc i, est crée à l'aide de la clé privé. Les blocs et les tags sont stockés chez le fournisseur et l'utilisateur peut les supprimer localement. Pour vérifier la possession des données, l'utilisateur envoie une liste de c numéros de blocs, sélectionnés aléatoirement, la clé publique, et un nombre aléatoire r au fournisseur. Le fournisseur doit ensuite récupérer les c blocs correspondants et en utilisant la clé publique calculer un tag pour chacun, la valeur du hash de r et envoyer le résultat à l'utilisateur. À la réception de la preuve, l'utilisateur utilise sa clé privée afin de conclure de la bonne possession des données ou non. Ce protocole évite la consommation de bande passante tout en empêchant le fournisseur de mentir sur la possession des données, mais nécessite que le serveur soit actif dans le calcul de la preuve, ce qui rend sa mise en place difficile dans le cadre des méthodes étudiées.

Tableau 3.1 – Classification des méthodes d'estimation de la localisation à partir de points de repère

	Corrélation @PA/Pos.	**Points de repère**			**Collecte de mesures**		**PDP**
		Type	**Échelle de distribution**	**Sélection**	**Métriques**	**Protocole de sonde**	
Biswal et al. [38]	Oui	Actif	USA	Aucune	RTT, HC, BW	ICMP	Aucun
Ries et al. [39]	Oui	Actif	Mondiale	Aucune	RTT	ICMP	Aucun
Fotouhi et al. [40]	Oui	Actif	Mondiale	Pré-vérification	RTT	ICMP	Aucun
Jaiswal et Kumar [41]	Non	Actif	USA	Pré-vérification	RTT	ICMP, HTTP(S)	Aucun
Benson et al. [42]	Non	Actif	USA	Aucune	RTT	HTTP(S)	Aucun
Gondree et Peterson [43]	Non	Actif	USA	Aucune	RTT	HTTP(S)	MAC
Watson et al. [44]	Non	Actif	USA et Europe	Pré-vérification	RTT	HTTP(S)	MAC
Eskandari et al. [45]	Non	Passif	Mondiale	Pré-apprentissage	RTT	HTTP(S)	Aucun

	Apprentissage automatique			
	Coordination de l'entrainement	**Estimation de la distance**	**Inférence de la localisation**	**Granularité de la localisation**
Biswal et al. [38]	Centralisé	Aucune	Classif. Naïve Bayésienne	Conté
Ries et al. [39]	Centralisé	Coordonnées virtuelles	Classif. Instance-Based	Pays
Fotouhi et al. [40]	Décentralisé	Bestline	Multilatération	Variable (Zone)
Jaiswal et Kumar [41]	Décentralisé	Ratio délai sur distance	Multilatération	Coordonées GPS
Benson et al. [42]	Centralisé	Régression linéaire	Regroupement hierarchique	Ville
Gondree et Peterson [43]	Décentralisé	Bestline	Multilatération	Variable (Zone)
Watson et al. [44]	Décentralisé	Régression linéaire	Multilatération	Variable (Zone)
Eskandari et al. [45]	Centralisé	Régression polynomiale	Multilatération	Coordonées GPS

3.4 Synthèse des résultats expérimentaux

Chaque méthode présentée dans ce chapitre propose ses propres résultats expérimentaux afin de mettre en valeur ses performances. Afin de pouvoir évaluer et comparer ces méthodes, les contextes expérimentaux et les résultats revendiqués sont présentés ici. L'ensemble des résultats expérimentaux rapportés par les auteurs des méthodes sont disponibles dans le tableau 3.2.

3.4.1 Contexte expérimental

Le contexte expérimental peut être divisé en 4 critères :

— Le nombre de points de repère, c'est-à-dire combien de points de repères sont sélectionnés dans l'ensemble initial. Ce nombre varie bien sûr selon les méthodes, mais il est à remarquer que les méthodes à points de repères actifs utilisent un nombre « raisonnable » de points de repère, entre 9 et 89, alors que les méthodes à points de repère passifs peuvent se permettre d'utiliser de nombreux points de repère, jusqu'à 38 892 [45].

— Le service de points de repère ou autrement dit, le service fournissant les serveurs utilisés comme de points de repère. Cela peut être des nœuds Planetlab [51] ou des machines virtuelles pour des points de repères actifs, car il est nécessaire, dans ce cas, de les exploiter directement. Dans le cas de points de repères passifs, il est possible d'utiliser des sites web dont la position physique du serveur est connue afin de limiter les coûts.

— L'échelle de distribution des points de repère, qui est identique au critère de la section précédente, c'est-à-dire qu'elle représente la zone dans laquelle les points de repères sont déployés.

— La plateforme de stockage testée : différents fournisseurs de services Cloud existent et peuvent se comporter différemment, il était donc important de relever lequel était testé en conditions réelles. De plus, il est aussi possible d'utiliser un point de repère simulant le fournisseur, ce qui permet de tester différentes position géographiques.

— Les paramètres et conditions particulières des expérimentations. Plusieurs méthodes présentent différentes expérimentations, changeant un ou plusieurs paramètres ou conditions entre elles. Cela peut être les métriques prises en compte, les conditions de sélections de points de repère, etc.

3.4.2 Résultats

Il y a trois critères de résultat :

— Le taux de succès, qui représente le pourcentage de résultats dont l'erreur de distance est égale ou inférieure à celle reportée. À une erreur de distance égale, un taux de succès plus important indique une meilleure solution.

— La granularité, qui est identique au critère de la section précédente, c'est-à-dire qu'elle représente la finesse du résultat retourné à l'utilisateur.

— L'erreur de distance qui est la distance entre le résultat retourné à l'utilisateur et le vrai lieu de stockage (ou stockage « simulé »). Le calcul peut se faire de différentes manières, selon la granularité du résultat. Par exemple, dans le cas de coordonnées GPS il s'agit de la distance entre les coordonnées retournées et les coordonnées du lieu de stockage. Dans le cas où le résultat est interprété avant d'être retourné, comme un conté, la distance est celle entre la frontière la plus proche du lieu de stockage et les coordonnées du lieu de stockage. Dans d'autres cas, cette erreur de distance n'est pas calculée et peut être considérée comme nulle. À un taux de succès égal, une erreur de distance plus faible indique une meilleure solution.

Les indicateurs « taux de succès » et « erreur de distance » sont liés et généralement, à moins de changer le contexte expérimental, quand l'un s'améliore l'autre se détériore. C'est à l'utilisateur de choisir un compromis entre les deux.

Tableau 3.2 – Contextes des expérimentations et résultats pour les méthodes d'estimation de la localisation à partir de points de repère

	Contexte expérimental					Résultats		
	Nombre de PdR	**Service de PdR**	**Échelle de distribution**	**Plateforme de stockage**	**Paramètres et conditions particulières**	**Taux de succès**	**Granularité**	**Erreur de distance (km)**
Biswal et al. [38]	67	Planetlab	USA	Amazon	Moy. et Std. du RTT et HC	95.00%	Conté	1,6
						100%	Conté	64
				Rackspace		99,99%	Conté	1,6
						100%	Conté	2 253
				Google		99,99%	Conté	1,6
						100%	Conté	1 609
				Amazon	Identiques + BW	100%	Conté	1,6
Ries et al. [39]	80 - 89	Planetlab	Mondiale	PdR	Phoenix	90%	Pays	
					Pharos	80%	Pays	
					Vivaldi	65%	Pays	
					Phoenix + proxy	50%	Pays	
					Vivaldi + proxy	40%	Pays	
					Pharos + proxy	30%	Pays	
Fotouhi et al. [40]	9	VM Amazon	Mondiale	Google		100%	Variable (zone)	240
Jaiswal et Kumar [41]	60	Planetlab	USA	PdR		100%	Coord. GPS	88,5
				Amazon		100%	Coord. GPS	112,7
Benson et al. [42]	36	Planetlab	USA	PdR		100%	Ville	441,6
Gondree et Peterson [43]	50	Planetlab	USA	PdR		50%	171 819 km^2	166
						90%	1 960 510 km^2	626
				Amazon	RTT min.	100%	11 175 km^2	
					RTT médian	100%	243 791 km^2	
Watson et al. [44]	28	Planetlab	USA et Europe	PdR	Europe	50%	Coord. centre	800
						75%	Coord. centre	1 000
					Amerique du Nord	50%	Coord. centre	1 000
						75%	Coord. centre	1 200
Eskandari et al. [45]	38 892	Sites web	Mondiale	PdR	PdR <100 km	100%	Coord. GPS	100
					PdR <1 000 km	100%	Coord. GPS	600

3.5 Limites et problèmes de ces méthodes

Ces méthodes fournissent la possibilité d'offrir à l'utilisateur une estimation sur la position de ses données dont la précision dépend des techniques employées par la méthode et de la qualité des mesures, sous la condition qu'elles soient correctement implémentées. Cependant même dans ce cas elles présentent des limites et soulèvent certains problèmes, qui seront abordés dans la suite de cette section.

3.5.1 Aucune garantie sur la position des données

Contrairement aux méthodes présentées dans le chapitre précédent, celles présentées dans celui-ci n'apportent aucune garantie sur la localisation des données. En effet, ces méthodes sont soumises à une marge d'erreur, qui dépend à la fois de paramètres contrôlables mais aussi de paramètres non-contrôlables. Les paramètres contrôlables, comme les conditions expérimentales (le nombre de points de repère, leur répartition, les métriques, etc.), permettent de réduire l'erreur de localisation de ces méthodes, quand ils sont choisis de manière cohérente avec la cible à tester.

Cependant, il reste des paramètres incontrôlables qui ont un impact important sur le fonctionnement du processus de vérification. Par exemple, la qualité des mesures dépend du réseau, et l'utilisateur n'en a pas le contrôle total. Il n'a donc aucun mécanisme lui permettant d'évaluer à la volée la qualité des mesures réalisées. De plus, si les mesures varient fortement entre l'étape d'apprentissage et l'étape de vérification, elles peuvent mener à des résultats incorrects (faux positifs ou faux négatifs), sans que l'utilisateur ne puisse différencier le résultat incorrect d'une réelle localisation incorrecte.

Enfin, contrairement aux mécanismes mis en place par les méthodes présentées dans le chapitre précédent, les mécanismes présentés ici, même dans le cas théorique où la marge d'erreur serait nulle, n'ont aucune valeur légale et ne permettent pas à l'utilisateur de se retourner contre le fournisseur s'il détecte un lieu de stockage différent de celui défini par le SLA.

3.5.2 Compromission du processus de vérification

Un autre problème de ces méthodes est la compromission possible du processus de vérification. En effet, il est possible pour un fournisseur de mettre en place des mécanismes compromettant le processus de vérification [37]. Certains de ces mécanismes ne concernent que des fournisseurs malicieux, qui souhai-

teraient déplacer le lieu de stockage afin de réduire les coûts sans en informer l'utilisateur. Néanmoins, des mécanismes sont applicables par tous types de fournisseurs sur le principe de la sécurité. Par exemple :

— Le seul moyen d'accéder aux données depuis l'extérieur est d'utiliser le point d'accès fourni, généralement grâce au protocole HTTP(S). Un fournisseur peut donc bloquer tous les autres protocole, et particulièrement ICMP. De plus le protocole ICMP est un protocole de choix des attaques DDoS [52], il n'est donc pas étrange qu'il soit bloqué.

— Bloquer les points de repère, qui ont un comportement « suspect » du point de vue du fournisseur. Plusieurs dizaines d'accès simultanés au même serveur, répétés régulièrement, afin de détecter un changement de lieu de stockage éventuel, peuvent être considérés comme « suspects » par le fournisseur. De plus ces accès s'effectuent en général sans modifier les données concernées, soit en effectuant un « ping », soit en accédant à des données de manière aléatoire, ce qui peut faciliter la détection par le fournisseur.

— Utiliser un proxy, souvent pour répartir la charge [53, 54], a aussi pour effet de masquer la topologie interne du réseau. Les méthodes ne faisant pas d'accès aux données estiment ainsi la position du proxy, qui peut se trouver dans un lieu différent de celui du stockage des données.

3.5.3 Absence de cadre unifié

Le dernier problème de ces méthodes n'est pas intrinsèquement lié aux méthodes en elles-mêmes mais plutôt à la présentation qui en est faite. Dans la littérature, chaque auteur souhaite valoriser ses résultats, parfois en les comparant aux résultats d'autres méthodes pour montrer les améliorations réalisées. Cependant, les hypothèses de départ et les conditions expérimentales sont différentes entre chaque méthode, il n'y a pas de cadre unifié pour la conception ni d'indicateurs unifiés pour les résultats, comme il est possible de remarquer dans le tableau 3.2.

Il n'est donc facile pour un utilisateur de choisir une méthode qui lui convienne car les résultats présentés sont issus de contextes différents, difficilement comparables.

Deuxième partie

Contributions

Chapitre 4

Cadre générique pour l'uniformisation des algorithmes de localisation des données

Sommaire

4.1 Introduction

Comme évoqué dans le chapitre 3, les différentes méthodes n'ont pas de cadre unifié. En effet chaque méthode, apporte sa propre notation, définit différentes hypothèses et conditions expérimentales, et utilise différents scores. Toutes ces différences ne facilitent pas la conception de nouvelles méthodes. Pour la même raison, elles rendent aussi le choix d'une méthode difficile pour un utilisateur. En effet, il n'y pas de cadre permettant de les comparer.

Bien qu'elles utilisent différents algorithmes, notamment pour l'apprentissage, il est possible de remarquer que leur structure et leur fonctionnement général est similaire. En effet, toute méthode fonctionne de la manière suivante :

— Des définitions initiales, qui comprennent l'ensemble initial des points de repère, le fournisseur c'est-à-dire la cible, qui sera testé et le modèle qui sera entrainé lors de l'apprentissage.

— Une étape d'apprentissage durant laquelle les points de repère interagissent entre-eux afin de définir les paramètres du modèle choisi, avec optionnellement une sélection des points de repère.

— Une étape de vérification durant laquelle les points de repère interagissent avec la cible, afin d'estimer la position géographique de cette dernière en utilisant le modèle entrainé pendant l'apprentissage.

— Une évaluation de la méthode qui prend en compte les propriétés connues de la cible testée (comme sa position géographique) et le résultat de l'estimation pour calculer un ou plusieurs scores servant à évaluer l'efficacité de la méthode.

Il est proposé dans ce chapitre une uniformisation de la notation utilisée par les algorithmes permettant l'estimation de la localisation des données à l'aide de points de repère. Pour cela, les différentes étapes d'une méthode seront définies et détaillées, à savoir les définitions initiales des paramètres de la méthode, le fonctionnement de l'étape d'apprentissage et celui de l'étape de vérification, et finalement le fonctionnement de l'étape d'évaluation. Pour l'évaluation, des scores prenant en compte la qualité du résultat seront proposés.

Le cadre proposé se concentre sur les méthodes entrainant une fonction liant les mesures réseau entre deux points à la distance physique qui les sépare lors de l'apprentissage et utilisant le procédé de multilatération pour inférer la position de la cible lors de la vérification. Cependant, il est assez générique pour être adaptable à l'ensemble des méthodes avec quelques hypothèses.

Le fonctionnement général, en blocs de construction, est repris par la fi-

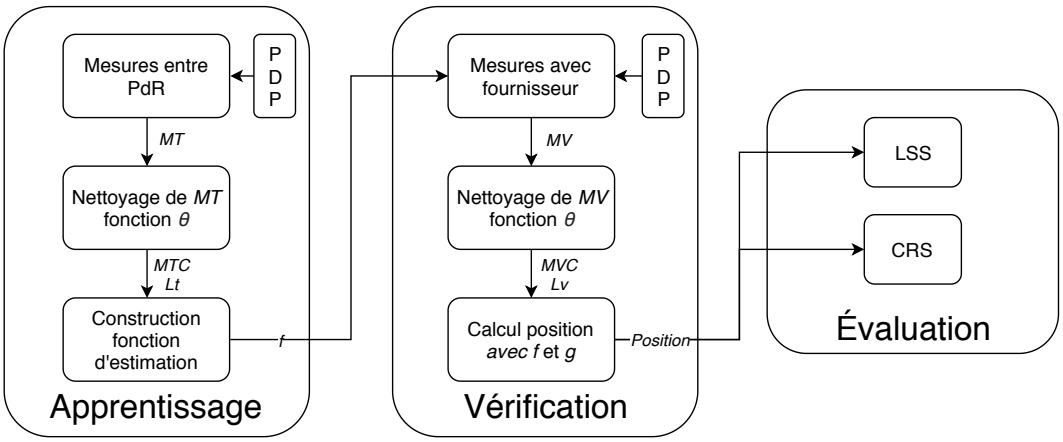


Figure 4.1 – Blocs de construction des algorithmes de localisation

gure 4.1.

4.2 Définitions initiales

La première notation à définir est l'ensemble des points de repères qui est noté L (pour Landmarks) et représente les différents points de repères sélectionnés pour la mise en place de la méthode. Par exemple :

$$L = \{l_1, l_2, ..., l_n\}$$

De plus, il est considéré une fonction Loc qui permet d'associer sa position à un point de repère donné. La position peut être sous forme d'étiquette ou bien sous forme de coordonnées, selon la définition de la fonction. Par exemple :

$$Loc(l_1) = Toulouse$$
$$\text{ou } Loc(l_1) = (43,604652; 1,444209)$$

L'ensemble des cibles testées est noté T (pour Target) et est aussi à définir. Les cibles représentent les endroits dans lesquels les données sont supposées être stockées. Dans le cas général l'ensemble T est composé de g différentes cibles :

$$T = \{t_1, t_2, ..., t_g\}$$

Dans le cas où il n'y a qu'une seule cible, l'ensemble est constitué d'un seul élément. Par exemple, en supposant que les données peuvent être sockées à

Toulouse :

$$T = \{t_1\}$$
$$Loc(t_1) = Toulouse$$

Une mesure réseau entre deux points est notée $m^{orig,dest}$ avec $orig$ le point initial la mesure et $dest$ le point mesuré. Si plusieurs mesures sont réalisées entre deux points, elles sont indicées. Par exemple, la i^{e} mesure entre l_1 et l_2 est notée :

$$m_i^{l_1,l_2}$$

Toutes ces mesures sont réunies au sein d'un ensemble M (pour Measures), ce qui permet de définir les ensembles MT utiles pour l'apprentissage et MV utiles pour la vérification :

$$\text{Mesures entre points de repères} = MT = \{m_i^{orig,dest} \in M \mid dest \in L\}$$
$$\text{Mesures vers une cible (Toulouse)} = MV = \{m_i^{orig,dest} \in M \mid dest = Toulouse\}$$

Les ensembles MT (pour Measures for Training) et MV (pour Measures for Verification) peuvent être construits « en ligne », c'est-à-dire en même temps que l'exécution de la méthode, comme présenté dans ce chapitre. L'autre manière de les construire, « hors ligne », consiste à d'abord collecter l'ensemble M puis de le séparer en MT et MV lors de l'utilisation de la méthode.

Une fonction $mesure$ est aussi définie, elle permet de collecter des mesures réseaux entre deux points. Ainsi :

$$mesure(l_1, l_2) = (13; 4)$$

Dans cet exemple la fonction $mesure$ collecte le couple $(RTT; Sauts)$ entre l_1 qui initie la mesure et l_2 le nœud sondé. Les métriques collectées sont définies par l'utilisateur et le fonctionnement de la fonction l'est aussi. Le fonctionnement correspond à la façon de collecter les mesures et le protocole utilisé, c'est-à-dire si la fonction réalise un simple « ping » pour récolter le RTT avec ICMP, ou bien utilise le point d'accès du fournisseur avec HTTP(S). Dans le cas où HTTP(S) est utilisé, une PDP peut être incluse ou non, ce mécanisme étant optionnel.

De plus, on suppose l'existence une fonction A, calculant l'aire d'une figure géométrique simple ou composée permettant par la suite de calculer les scores. On suppose aussi l'existence d'une fonction C telle que $C(x, r)$ désigne un

cercle de rayon r et de centre x. Par exemple :

$$A(C((0,0),1)) = \pi$$

$$A(C((0,0),1) \cap C((0,1),1) = \frac{2\pi}{3} - \frac{\sqrt{3}}{2}$$

On définit aussi préalablement un type de fonction d'estimation pour la fonction f, tel que fonction polynomiale, fonction exponentielle, etc. Notons qu'à ce stade, les paramètres de cette fonction ne sont pas définis. Seule la manière dont ils seront obtenus est déterminée. Cette fonction prend comme paramètre une mesure réalisée entre deux nœuds et retourne un résultat intermédiaire. Généralement, ce résultat est l'estimation de la distance qui sépare les deux nœuds. Cette fonction peut être propre à chaque point de repère, c'est-à-dire que chaque point de repère aura des coefficients différents, mais toujours un même type de fonction. Elle peut aussi être globale, et chaque point de repère utilise la même fonction. Dans les deux cas, elles sont indicées par point de repère, par exemple pour une fonction polynomiale de degré 3 :

$$\begin{aligned} f_{l_1}(m^{l_1,l_2}) = & \, a_3^{l_1} \times (m^{l_1,l_2})^3 \\ & + a_2^{l_1} \times (m^{l_1,l_2})^2 \\ & + a_1^{l_1} \times m^{l_1,l_2} \\ & + a_0^{l_1} \end{aligned}$$

Les coefficients a_{l_1}, b_{l_1}, c_{l_1}, d_{l_1} sont à déterminer pendant l'apprentissage.

Comme chaque point de repère possède une fonction associée, f_{l_i}, les paramètres de cette fonction seront toujours de la forme $m_i^{l_i,dest}$ avec $dest \in T \cup L$.

Finalement, il existe une fonction g qui prend en paramètre les résultats des f_{l_i} des différents points de repère, directement, ou après transformation, afin d'estimer la localisation. Une transformation possible est d'associer à chaque résultat intermédiaire représentant la distance, un cercle centré autour de la position du point de repère initiant la mesure. Le type de résultat de cette fonction dépend directement de la granularité de la méthode.

$$t_i \in T, \forall l_j \in L, \widehat{d_{l_j,t_i}} = f_{l_j}(m^{l_j,t_i})$$
$$\widehat{t_i} = g(\, C(Loc(l_1), \widehat{d_{l_1,t_i}}),$$
$$C(Loc(l_2), \widehat{d_{l_2,t_i}}),$$
$$...,$$
$$C(Loc(l_n), \widehat{d_{l_j,t_n}}))$$

4.3 Apprentissage

L'apprentissage consiste à déterminer les paramètres de la fonction f à l'aide de mesures réseau considérées comme caractérisant le réseau de manière statistique. Pour le réaliser il y a trois étapes :

4.3.1 Mesures entre points de repère

La première étape de l'apprentissage est de construire l'ensemble MT avec des mesures inter-points de repère. Chaque point de repère de l'ensemble initial va interroger tous les autres, ainsi que lui-même, pour récolter des mesures réseau m à l'aide de la fonction $mesure$. Cette fonction va donc réaliser les mesures depuis le point de repère représenté par son premier paramètre, vers celui représenté par son second paramètre. Par exemple, une mesure entre l_1 et l_2 est définie $m^{l_1,l_2} = mesure(l_1, l_2)$. Ainsi, chaque mesure m comprend les mêmes types de données. Par exemple, si les RTTs et les nombres de sauts sont considérés, les différentes mesures seront toujours de la forme $(RTT; Sauts)$: $(m_1, m_2, ..., m_K) = ((3; 4), (13; 5), ..., (8; 4))$.

Si une PDP est mise en place, elle est incluse dans la définition de la fonction $mesure$, c'est à dire que la fonction $mesure$ doit être capable d'indiquer si la PDP est valide ou non. En effet, l'utilisation de PDP peut être vue comme une façon de mesurer particulière. Pour cela, nous proposons un retour de la fonction $mesure$ indiquant si la PDP a réussi ou non. Par exemple :

$$mesure(l_1, l_2) = \begin{cases} \text{Les valeurs mesurées si la PDP est valide} \\ \text{Des valeurs hors domaine mesurable sinon} \end{cases}$$

Comme certaines méthodes utilisent des valeurs brutes, d'autres le RTT

minimum, et d'autres la moyenne, afin de caractériser au mieux le réseau, le procédé d'interrogation est réalisé K fois. C'est-à-dire que l'ensemble MT est défini de la manière suivante :

$$MT = \{m_k^{l_i,l_j} \mid \forall l_i, l_j \in L \times L, \\ \forall k \in \{1, 2, ..., K\}, m_k^{l_i,l_j} = mesure(l_i, l_j)\}$$

On peut aussi noter l'ensemble MT^{l_i}, l'ensemble des mesures réalisées depuis le point de repère l_i vers tous les autres points de repère :

$$MT^{l_i} = \{m^{orig,dest} \in MT \mid orig = l_i\}$$

4.3.2 Sélection des points de repère et nettoyage de MT

Une fois l'ensemble MT construit, c'est-à-dire les mesures effectuées, il est possible que certaines mesures présentent des anomalies et il n'est pas souhaitable de les conserver. Il faut donc nettoyer l'ensemble MT de ces mesures anormales. Pour cela on suppose l'existence d'une fonction θ prenant en paramètre l'ensemble MT initial et retournant l'ensemble MT nettoyé. La définition exacte de la fonction θ n'est pas décrite mais elle permet de sélectionner les valeurs considérées comme fiables et représentatives. Bien sûr, il ne s'agit pas de sélectionner arbitrairement les mesures à supprimer afin de biaiser la méthode, pour qu'elle ait de meilleurs ou de pires résultats. Les mesures à supprimer le sont parce qu'elles ne respectent pas certains critères et ne caractérisent pas, ou mal, le réseau.

Voici par exemple, plusieurs ensembles de valeurs que peut exclure la fonction θ :

— Une première raison pour retirer des mesures est que certaines valeurs sont normalement impossibles à recueillir. Par exemple les RTTs négatifs sont par définition impossibles, ces valeurs sont donc à éliminer de MT. Ces valeurs peuvent provenir d'un problème au niveau de la fonction de collecte qui ne se serait pas exécutée normalement. De même, les RTT ne peuvent pas être nuls, car même s'ils peuvent rentrer dans la définition d'un RTT, ils sont en pratique impossibles à recueillir même s'ils sont proviennent depuis un point de repère vers lui-même (un temps de calcul non nul est présent). Les nombres de sauts sont aussi non négatifs ou nuls (il y a toujours au moins un saut). De plus, le champ TTL de l'entête IPv4 est codé sur 8 bits [55], ce champ est utilisé pour déterminer le nombre de saut, il ne peut donc pas être supérieur à 255. On peut ainsi, par exemple,

définir un ensemble V_{imp} qui comprend ces valeurs, en considérant les mesures comme des couples $(RTT; Sauts)$:

$$V_{imp} = (\mathbb{R}_{\leq 0} \times \mathbb{Z}) \cup (\mathbb{R} \times (\mathbb{Z}_{\leq 0} \cup \{n \mid n \in \mathbb{N} \wedge n > 255\}))$$

Bien sûr d'autres valeurs peuvent être ajoutées à cet ensemble selon les métriques considérées. L'ensemble MT devient donc MTP (pour Measures for Training : Possible) :

$$MTP = \{m^{orig,dest} \in MT \mid m^{orig,dest} \notin V_{imp}\}$$

— Une autre raison est que certaines valeurs sont incohérentes. Tout en restant dans l'ensemble des valeurs possibles, les valeurs minimales de RTT mesurables entre deux nœuds sont limitées. En effet, en considérant un réseau câblé à la fibre optique de bout en bout, la vitesse à laquelle l'information traverse physiquement le réseau pour réaliser l'aller et le retour nécessaire au calcul du RTT est limité par la vitesse de la lumière dans la fibre, c'est-à-dire environ $\frac{2}{3}$ de la vitesse de la lumière dans le vide [56] par traversée. Comme les positions des points de repère sont connues, la distance à vol d'oiseau qui les sépare l'est aussi. De plus, la distance réelle parcourue par l'information est plus élevée et le réseau n'est probablement pas entièrement composé de fibre optique. Il serait possible d'affiner cette vitesse de traversée avec une plus grande connaissance de la topologie. Des valeurs de RTT qui ne respecteraient pas cette contrainte et impliqueraient une vitesse de traversée plus rapide sont aussi à éliminer. On suppose donc qu'il existe un prédicat inc capable de déterminer si une mesure m est incohérente en fonction de son origine et sa destination. Ainsi, l'ensemble des mesures valides devient MTV (pour Measures for Training : Valid) :

$$MTV = \{m^{orig,dest} \in MT \mid !inc(m^{orig,dest})\}$$

— Finalement, certains points de repères peuvent ne pas envoyer ou répondre aux requêtes, ou produire des réponses incohérentes ou impossibles, à cause d'un problème du point de repère ou du réseau. Même si cela arrive par intermittence, il suffit qu'un certain seuil d'erreurs soit atteint pour que les mesures provenant ou à destination de ces points de repère ne soient pas intéressantes pour caractériser le réseau. Il n'est pas souhaitable de conserver de tels points de repère. En supposant l'exis-

tence d'un prédicat $fiable$, déterminant si un point de repère réalise un nombre convenable de mesures, l'ensemble des points de repères devient Lt défini tel que :

$$Lt = \{l \in L \mid fiable(l)\}$$

Le prédicat $fiable$ peut bien sûr déclarer que tous les points de repère sont fiables s'ils répondent tous correctement. Ainsi dans certains cas $Lt = L$.

Il est ensuite possible de construire l'ensemble des mesures provenant de points de repères fiables, d'où une deuxième façon d'écrire l'ensemble des valeurs fiables MTR (pour Measures for Training : Reliable) :

$$MTR = \{m^{orig,dest} \in MT \mid orig, dest \in Lt \times Lt\}$$

Finalement l'ensemble MTC ((pour Measures for Training : Cleaned), représentant l'ensemble nettoyé peut se définir comme :

$$\begin{aligned} MTC = \theta(MT) = \{m^{orig,dest} \in MT \mid & \, m^{orig,dest} \notin V_{imp} \\ & \wedge !inc(m^{orig,dest}) \\ & \wedge orig, dest \in Lt \times Lt\} \end{aligned}$$

4.3.3 Calcul des paramètres de la fonction d'estimation

Le but de l'apprentissage est de définir les paramètres de la fonction ou des fonctions f à partir des mesures de l'ensemble MTC. Généralement cette fonction estime une distance physique en fonction d'une mesure réseau, par exemple :

$$p, q \in Lt \times Lt$$
$$Loc(p) = Toulouse, Loc(q) = Marseille$$
$$f_p(m^{p,q}) = \widehat{d_{p,q}} = \widehat{d_{Toulouse,Marseille}} = 400 \text{ (km)}$$

En considérant le cas où chaque point de repère à sa propre fonction, elle se construit en sélectionnant pour chaque point de repère les mesures qu'il a effectuées, soit l'ensemble MTC^{l_i}. Pour chaque mesure de cet ensemble, il est

possible de retrouver la cible l_j. En connaissant l_j, la distance entre l_i et l_j est calculable par une fonction *distance* (en fonction des positions des points de repère), il est donc possible d'associer à chaque mesure, la distance entre l'origine et la cible :

$$\begin{aligned}\forall l_i \in Lt, N_{l_i} = &\{(m^{l_i,q}; d_{l_i,q}) \mid m^{l_i,q} \in MTC^{l_i} \\ &\wedge d_{l_i,q} = distance(Loc(l_i), Loc(q))\}\end{aligned}$$

Cet ensemble N_{l_i}, composé d'éléments $(échantillon, label)$ est ensuite utilisé pour déterminer les coefficients de la fonction f_{l_i} selon la méthode définie au départ, par régression linéaire ou polynomiale, par bestline, etc. En supposant une régression linéaire :

$$linreg(N_{l_i}) = (a_{l_i}, b_{l_i})$$

Quand la fonction f est calculée globalement, le fonctionnement de l'estimation des paramètres est identique mais l'ensemble N n'est pas associé à un point de repère particulier, l'ensemble des origines est considérée en même temps lors de la détermination des paramètres.

4.4 Vérification

La vérification consiste à estimer la position physique de l'ensemble des cibles en utilisant les fonctions f construites lors de l'apprentissage et la fonction g définie au départ. Pour cela il y a aussi trois étapes, identiques à celles de la section 4.3 mais avec le fournisseur pour cible et non plus les points de repère.

4.4.1 Mesures entre points de repère et fournisseur

La première étape de la vérification est de construire l'ensemble MV avec des mesures depuis les points de repère vers l'ensemble des cibles. Chaque point de repère de l'ensemble d'apprentissage va interroger toutes les cibles pour récolter des mesures réseau m à l'aide de la même fonction *mesure*, utilisée lors de l'apprentissage. Comme lors de l'apprentissage, chaque mesure m comprend les mêmes types de données.

Comme certaines méthodes utilisent les valeurs brutes, d'autres le RTT

minimum, et d'autres la moyenne, afin de caractériser au mieux le réseau, le procédé d'interrogation est réalisé K fois. C'est-à-dire que l'ensemble MV est défini de la manière suivante :

$$MV = \{m_k^{l_i,t_j} \mid \forall l_i, t_j \in L \times T, \\ \forall k \in \{1, 2, ..., K\}, m_k^{l_i,t_j} = mesure(l_i, t_j)\}$$

On peut aussi noter l'ensemble MV^{t_i}, l'ensemble des mesures réalisées vers une cible t_i :

$$MV^{t_i} = \{m^{orig,dest} \in MV \mid dest = t_i\}$$

4.4.2 Sélection des points de repère et nettoyage de MV

De la même manière que lors de l'apprentissage, les valeurs des mesures récoltées peuvent présenter des anomalies. L'ensemble V_{imp} et le prédicat inc, considérés par la fonction θ sont toujours valides et leur définition reste inchangée.

Cependant, la définition précédente du prédicat $fiable$ n'est pas exhaustive et peut être modifiée. En effet, en plus d'éliminer les points de repère n'effectuant pas assez de mesures sur les K à réaliser, il est possible d'exclure ceux dont les valeurs sont trop élevées, indiquant leur éloignement de la cible, et qui ne contribueront donc pas à améliorer la qualité du résultat. L'ensemble des points de repères utilisés par la vérification devient donc :

$$Lv = \{l \in Lt \mid fiable(l)\}$$

Comme lors de l'apprentissage, l'ensemble des points de repères Lv peut être identique à Lt.

Finalement, l'ensemble MVC (pour Measures for Verification : Cleaned), la version nettoyée de MV, est défini tel comme lors de l'apprentissage :

$$MVC = \theta(MV) = \{m^{orig,dest} \in MV \mid m^{orig,dest} \notin V_{imp} \\ \wedge !inc(m^{orig,dest}) \\ \wedge orig, dest \in Lv \times Lv\}$$

4.4.3 Estimation de la localisation

Une fois les mesures collectées et triées, la partie d'estimation de la localisation consiste, pour chaque cible, à utiliser ces mesures et les différentes

fonctions f afin d'estimer le résultat intermédiaire pour chaque point de repère, c'est-à-dire, le plus souvent l'estimation de sa distance par rapport à la cible. Ce résultat peut être utilisé tel quel ou bien transformé. Par exemple, quand des distances sont estimées, un cercle autour du point de repère ayant pour rayon la distance estimée est utilisé à la place des distances brutes. L'ensemble des cercles $\mathscr{C}_i$ est donc construit à partir des distances estimées.

$$\forall t_i \in T, \mathscr{C}_i = \{C(pos(p), f_p(m^{p,t_i})) \mid m^{p,t_i} \in MVC \wedge p \in Lv\}$$

L'ensemble des résultats est ensuite utilisée par une fonction g, qui va inférer la position loc_{t_i} avec les résultats intermédiaires afin d'établir un consensus entre les points de repère. Cette fonction peut être une fonction de multilatération, qui prend l'ensemble des cercles en paramètre et retourne le consensus donné par leur intersection.

$$\forall t_i \in T, g(\mathscr{C}_i) = \widehat{loc_{t_i}}$$

4.5 Évaluation

L'étape d'évaluation est cruciale. Lors de cette étape le lieu de stockage est connu, ainsi il est important de choisir des indicateurs pertinents pour comparer l'estimation par la méthode et la réalité connue. En effet, il est essentiel de pouvoir apprécier la qualité de la méthode (son taux de succès) et la qualité des résultats (leur précision) afin de la valoriser. Dans la littérature, chaque méthode propose une manière d'évaluer les méthodes, parfois avec des indicateurs calculés de façon spécifique. Pour palier à cela, deux scores calculables pour l'ensemble des méthodes sont proposés [57]. Le premier permet d'évaluer la qualité de la méthode, le suivant la qualité des résultats.

Afin de calculer ces scores, il est associé à chaque cible t_i une zone loc_{t_i}, représentée par une figure géométrique, au sein de laquelle le stockage des données est considéré comme autorisé. La taille et la forme de cette zone sont choisies afin de correspondre à la granularité souhaitée. Par exemple, si la cible est $Toulouse$ et la zone acceptée est un cercle de 10 km autour de cette ville, alors $loc_{Toulouse}$ est défini comme :

$$loc_{Toulouse} = C(pos(Toulouse), 10)$$

4.5.1 Score de succès de la localisation

Le premier score, le Score de succès de la localisation (LSS ou Location Success Score), permet de vérifier l'existence d'une intersection entre le consensus $\widehat{loc_{t_i}}$ et la cible loc_{t_i}. En effet, une méthode dont le LSS serait faible ne trouverait pas la zone de stockage. Ce score permet donc d'attester de la qualité de la méthode en elle-même en proposant un score se rapprochant d'un taux de succès.

Pour chaque cible t_i le LSS_i est calculé :

$$\forall t_i \in T, LSS_i(loc_{t_i}, \widehat{loc_{t_i}}) = \begin{cases} 1, & \text{si } loc_{t_i} \cap \widehat{loc_{t_i}} \neq \emptyset \\ 0, & \text{sinon} \end{cases}$$

Les différents scores sont ensuite moyennés pour l'ensemble des cibles :

$$\frac{1}{|T|} \sum_{t_i \in T} LSS_i(loc_{t_i}, \widehat{loc_{t_i}})$$

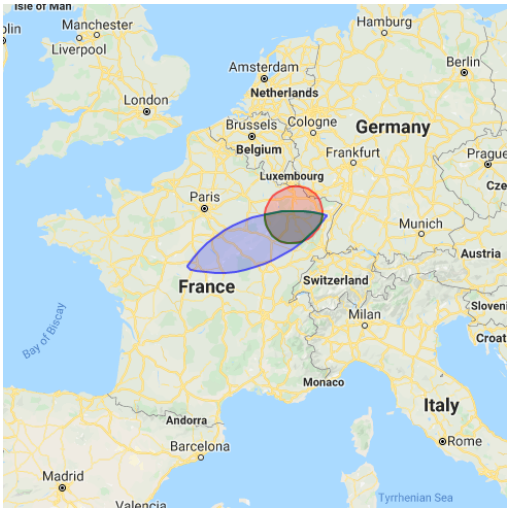
La figure 4.2 illustre le calcul de ce score. La zone bleue est un consensus et la zone rouge la cible. Lorsqu'il y a une intersection, comme sur la figure 4.2a, elle est représentée par une zone verte et le LSS vaut 1. À l'inverse, lorsqu'il n'y a pas d'intersection entre le consensus et la zone cible, comme sur la figure 4.2b le LSS vaut 0.

4.5.2 Score du ratio du consensus

Le deuxième score, le Score du ratio du consensus (CRS ou Consensus Ratio Score), permet de vérifier quelle proportion occupe l'intersection entre le consensus et la zone associée à la cible ($A(loc_{t_i} \cap \widehat{loc_{t_i}})$) par rapport à la taille totale du consensus ($A(\widehat{loc_{t_i}})$). En effet, une méthode dont le CRS serait trop bas, alors que le LSS est élevé, aurait tendance à créer des consensus beaucoup plus large que la zone associée aux cibles testées. En d'autre termes, si la cible est $Toulouse$, la méthode trouverait bien que les données sont stockées aux alentours de $Toulouse$ mais aussi de $Marseille$, $Lyon$ ou $Paris$. Ce score permet donc d'attester de la qualité des consensus fournis par la méthode car il permet d'estimer la précision du résultat fourni.

Pour chaque cible le CRS_i est calculé :

$$\forall t_i \in T, \widehat{loc_{t_i}} : CRS_i(loc_{t_i}, \widehat{loc_{t_i}}) = \frac{A(loc_{t_i} \cap \widehat{loc_{t_i}})}{A(\widehat{loc_{t_i}})}$$



(a) Intersection entre consensus et cible

(b) Pas d'intersection entre consensus et cible

Figure 4.2 – Illustration des scores

Les différents scores sont ensuite moyennés pour l'ensemble des cibles :

$$\frac{1}{|T|} \sum_{t_i \in T} CRS_i(loc_{t_i}, \widehat{loc_{t_i}})$$

La figure 4.2 peut aussi illustrer le calcul de ce score. Lorsqu'il y a une intersection, comme sur la figure 4.2a, le CRS est calculé et vaut le ratio entre l'aire de la zone verte sur l'aire de la zone bleue. À l'inverse, lorsqu'il n'y a pas d'intersection entre le consensus et la zone cible, comme sur la figure 4.2b le CRS n'est pas calculé.

4.6 Cas des autres méthodes

Comme mentionné dans l'introduction de ce chapitre, notre cadre est fondé en généralisant à partir des méthodes qui :

— entrainent une fonction liant les mesures réseau entre deux nœuds à la distance physique qui les sépare ;

— utilisent le procédé de multilatération pour inférer de la position de la cible.

Dans la catégorie des méthodes d'estimation de la localisation dans le Cloud

à l'aide de points de repère, deux types de méthodes n'entrent pas totalement sous cette étiquette. Ce sont les méthodes utilisant des points de repère passifs [45] et celles utilisant uniquement une technique de classification [38].

4.6.1 Points de repères passifs

Il y a deux points qui peuvent être problématiques pour que les méthodes utilisant des points de repères passifs entrent dans le modèle proposé :

— Les mesures réalisées sont réalisées depuis la cible à tester vers les points de repères, c'est-à-dire qu'au lieu d'avoir des mesures m^{l_i,t_j} ce sont des mesures m^{t_j,l_i}. Cependant, ce n'est pas un problème car les mesures réalisées dans un sens ou dans l'autre sont très similaires, il est possible de remplacer les unes par les autres. Cette affirmation sera confirmée par les données présentées dans le chapitre 5. Ainsi, il est possible de considérer que les mesures m^{l_i,t_j} peuvent tout à fait remplacer les mesures m^{t_j,l_i} initialement prévues.

— Une hypothèse forte est faite par ces méthodes, tout du moins par [45]. Cette hypothèse est que la position de la cible est supposée connue et utilisée dans la phase d'apprentissage, afin de construire la fonction f. Dans le cadre de l'évaluation de la méthode, il est bien sur possible d'utiliser la position de la cible, parce que les cas d'évaluation sont choisis avec une position connue. Cependant, il est difficile de dire vouloir estimer une position tout en la supposant déjà connue a priori. La position trouvée n'aura aucune cohérence et les scores ne refléteront que la qualité des consensus. Au mieux, il sera possible de détecter un changement de lieu de stockage. Pour palier à cela, il est proposé de transformer les méthodes avec points de repères passifs en méthodes avec points de repères actifs, utilisant une coordination de l'entrainement centralisée car c'est le type de méthode qui se rapproche le plus.

4.6.2 Méthodes reposant sur la classification

Les méthodes utilisant uniquement une technique de classification utilisent directement les mesures réseau pour inférer directement de la localisation. Pour cela la fonction f devient une fonction identité. En effet, il n'y a pas de résultats intermédiaires, la fonction f est rendue obsolète. Cependant, les mesures collectées durant la phase d'apprentissage servent à construire la fonction g, la fonction de classification. De par leur nature, ces méthodes sont forcément

centralisées, la fonction g est commune et utilise ainsi l'ensemble des mesures collectées pour être définie.

Les autres éléments restent inchangés et ces simples modifications suffisent à intégrer ces méthodes dans le modèle.

Chapitre 5

Collecte des données : plateforme et pré-traitements

Sommaire

5.1 Introduction

Nous avons réalisé la collecte de deux jeux de données. Un au niveau national, en utilisant la plateforme Grid'5000 [58], et un au niveau mondial, en utilisant la plateforme Amazon AWS [59]. La raison de collecter à deux niveaux différents est que cela permet d'établir si l'échelle modifie les performances des algorithmes. Le but, à terme, de cette collecte est de pouvoir utiliser les données récoltées afin de comparer différents algorithmes de localisation.

Les jeux de données ont étés collectés de manière « hors ligne », c'est-à-dire, d'après la notation introduite par le chapitre 4 que l'ensemble M a d'abord été construit, puis les ensembles MT (et MTC) ainsi que MV (et MVC) en ont été extraits. Cela correspond respectivement aux étapes « Mesures entre points de repères » et « nettoyage de MT » de l'étape d'apprentissage ainsi qu'aux étapes « Mesures avec le fournisseur » et « nettoyage de MV » de l'étape de vérification. Nous avons choisi de réaliser une collecte hors-ligne pour pouvoir obtenir un jeu de données conséquent, pouvant être réutilisé et pouvant servir de base de comparaison pour les différents algorithmes utilisés pour la vérification. En effet, il est possible d'utiliser les mêmes valeurs pour différents algorithmes, réduisant la dépendance à la qualité des mesures. En d'autres termes, une mesure de « mauvaise qualité » ne sera pas la cause d'une performance médiocre, comparativement aux autres, d'un algorithme, et inversement pour une mesure de « bonne qualité ».

Les deux jeux de données incluent le RTT comme métrique, dans les deux cas, il a été mesuré avec un simple « ping » et non avec une méthode par accès aux fichiers stockés qui permettrait de mettre en place une PDP. La raison de ce choix est au niveau des limitations de la plateforme de collecte au niveau national. En effet, la plateforme Grid'5000 fonctionne par réservations de nœuds, il était donc impossible de mettre en place un environnement Cloud, avec toutes les ressources nécessaires, pendant plus de quelques heures sur l'ensemble de la plateforme. De plus, même s'il est possible de réaliser la collecte par intermittence, plusieurs nœuds, voire la totalité, étaient nécessaires en même temps, réduisant encore plus les moments où la collecte aurait pu être réalisée et nous empêchant ainsi d'avoir un jeu de données de taille raisonnable en un temps acceptable. Le compromis a été de pouvoir faire tourner notre processus de collecte sur les serveurs frontaux de Grid'5000, car ils ne nécessitaient pas de ressources particulières à part une utilisation limitée de l'accès réseau.

Ce chapitre explique, pour chaque jeu de données, le choix des plateformes de collecte et des points de repères sélectionnés. Les métriques et le fonctionnement de la collecte sont décrits ainsi que les jeux de données et les prétrai-

tements réalisés avant de pouvoir utiliser les données pour la comparaison des méthodes.

5.2 Collecte au niveau national

Pour établir le premier jeu de données dont les mesures sont réalisées au sein d'un environnement contrôlé et dont l'échelle est limitée, nous avons choisi de collecter des mesures au niveau national. Plusieurs critères permettent de considérer l'environnement fourni par Grid'5000 comme contrôlé. D'abord, la position des nœuds est garantie, ce qui est un critère important lorsque la localisation est la caractéristique à estimer. Ensuite, la topologie les caractéristiques du réseau de la plateforme et connue, les résultats obtenus peuvent ainsi être placé dans un contexte connu, ainsi les biais extérieurs sont connus ou peuvent être expliqués. Tester les différentes méthodes au sein d'un environnement contrôlé permet d'établir si elles sont en mesure d'estimer la localisation correctement et de comprendre les résultats expérimentaux, sans biais lié aux conditions expérimentales.

5.2.1 Choix et répartition des points de repère

Afin de répliquer le fonctionnement des différentes méthodes, il était nécessaire d'utiliser différents points de repère répartis au sein d'une zone géographique donnée.

Plateforme Grid'5000

Les mesures ont été réalisées grâce à la plateforme Grid'5000, une plateforme d'expérimentation répartie sur huit sites, principalement en France, (figure 5.1) et totalisant environ 800 nœuds. Les différents sites sont reliés entre eux par une connexion dédiée de 10Gb/s fournie par RENATER [60]. Chaque site propose le même environnement logiciel, permettant ainsi une homogénéité des configurations et excluant un biais dans les résultats obtenus. L'accès à Internet est limité par la plateforme, sauf pour certaines applications sur liste blanche, isolant ainsi le réseau de l'extérieur.

Pour utiliser les différents nœuds un système de réservation est mis en place. Ce système permet de limiter l'accès exclusif aux nœuds par une seule personne.

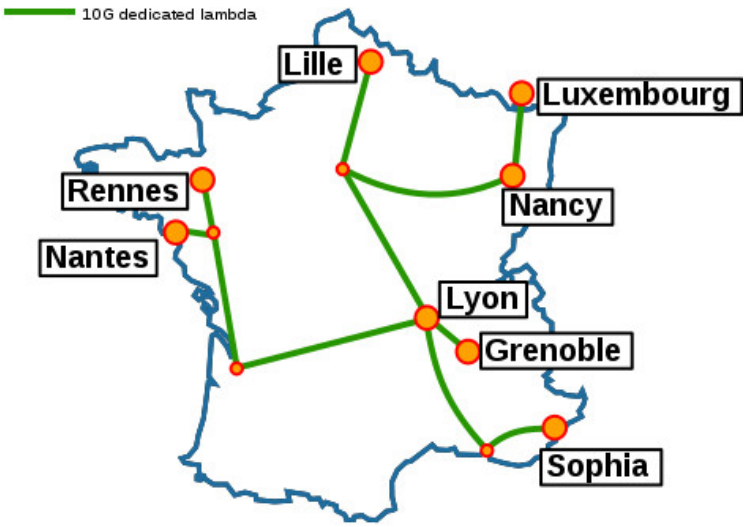


Figure 5.1 – Répartition des nœuds Grid'5000

Points de repères sélectionnés

Nous avons utilisé comme points de repère, l'ensemble des sites proposés par Grid'5000, c'est-à-dire : Grenoble, Lille, Luxembourg, Lyon, Nancy, Nantes, Rennes et Sophia. Notre processus de collecte ne nécessitait pas de ressources, à part un accès négligeable au réseau par rapport à la bande passante disponible, mais uniquement une exécution sur le long terme. Par conséquent, pour chaque site, nous avons pu avoir un accès au nœud frontal. Cet accès nous a permis d'éviter une réservation compliquée car l'ensemble des nœuds était nécessaire en même temps pour une longue période de temps. De plus, la collecte en est aussi facilitée car l'ensemble des adresses des nœuds à sonder étaient connues à l'avance, pouvant ainsi être fournies de manière statique.

Le nœud situé à Luxembourg, n'est effectivement pas en France, mais au Luxembourg. L'utilisation du terme « jeu de données national » pourrait ainsi être remise en question par sa présence. Cependant, la proximité du nœud avec la France combinée à l'utilisation d'un réseau privé, rend l'intégration de ce nœud au sein du jeu de données acceptable, car il se comporte a priori comme n'importe quel autre nœud et peut toujours être retiré a posteriori s'il se comporte différemment.

5.2.2 Métriques considérées

Deux métriques ont été considérées :

— Le RTT « simple » car c'est la métrique utilisée par toutes les méthodes, il était nécessaire de le collecter.

— Le nombre de sauts, bien qu'il ne soit utilisé uniquement par une seule méthode, était facile à collecter et peut être utile.

En plus de ces deux métriques, chaque mesure est accompagné d'un horodatage indiquant le moment de la collecte.

5.2.3 Fonctionnement de la collecte

La collecte a duré du 22 mai 2018 au 29 juin 2018, soit environ un mois avec un rythme d'une mesure toutes les cinq minutes. Un ensemble assez important de mesures était souhaité pour pouvoir caractériser le réseau.

Pour réaliser la collecte, un programme encapsulant la commande *traceroute* a été implémenté. Ce programme prend une liste de nœuds en entrée et les interroge en parallèle à une fréquence donnée, ici toutes les cinq minutes, puis enregistre les résultats. La liste des nœuds interrogés est l'ensemble des points de repère, c'est-à-dire que chaque nœud interrogeait les 7 autres nœuds et lui-même. L'interrogation se fait avec la commande *traceroute*. Cependant, certains routeurs bloquant les requêtes ICMP utilisées par la commande, nous utilisons un *traceroute* par TCP qui est généralement autorisé [61].

Ce programme a été lancé sur chacun des huit serveurs frontaux de Grid'5000 pendant le mois de la collecte.

5.3 Jeu de données national

5.3.1 Présentation du jeu de données initial

Le jeu de données comporte 614 244 mesures, collectées entre les 8 nœuds Grid'5000 servant de points de repère. La répartition des mesures selon l'origine et la destination sondée est détaillée au niveau du tableau 5.1. Plusieurs points peuvent être remarqués au sein de ce jeu de données.

D'abord, en observant la répartition des mesures, il peut être remarqué que deux nœuds, Lyon et Nancy, possèdent beaucoup moins de mesures que les autres. Ces deux nœuds étaient souvent redémarrés, par un administrateur de la plateforme Grid'5000, pour maintenance, interférant avec le processus de collecte.

De plus, les valeurs moyennes des RTT indiquent une similarité entre les couples $(origine; destination)$. C'est-à-dire que, par exemple, un RTT mesuré

Origine \ Dest.	Grenoble	Lille	Luxembourg	Lyon	Nancy	Nantes	Rennes	Sophia	Total
Grenoble	9 627	9 629	9 628	9 629	9 629	9 628	9 629	9 628	77 027
Lille	10 813	10 813	10 813	10 813	10 813	10 813	10 813	10 813	86 504
Luxembourg	10 839	10 839	10 839	10 839	10 839	10 839	10 839	10 839	86 712
Lyon	7 228	7 227	7 227	7 228	7 226	7 229	7 228	7 226	57 819
Nancy	7 258	7 258	7 258	7 258	7 258	7 258	7 258	7 258	58 064
Nantes	10 838	10 837	10 837	10 837	10 837	10 838	10 837	10 837	86 698
Rennes	9 337	9 340	9 338	9 339	9 339	9 339	9 337	9 338	74 707
Sophia	10 839	10 839	10 840	10 839	10 839	10 839	10 839	10 839	86 713
Total	76 779	76 782	76 780	76 782	76 780	76 783	76 780	76 778	614 244

Tableau 5.1 – Nombre de mesures collectées (jeu de données national brut)

depuis Grenoble vers Lille sera similaire à un RTT mesuré depuis Lille vers Grenoble. Ces RTT moyens ainsi que les écart-types selon l'origine et la destination sont détaillés au niveau des tableaux 5.2 et 5.3.

Il est aussi possible d'observer la répartition dans le temps des RTT (figure 5.2). Chaque couple de points de repère a un profil particulier de RTT selon les jours de la semaine. Par exemple, les RTTs collectés peuvent être similaires selon les jours de la semaine, comme entre Rennes et Grenoble (figure 5.2b), c'est-à-dire qu'ils sont répartis de la même manière entre les mêmes valeurs minimales et maximales pour chaque jour de la semaine. D'autres couples de points de repère, comme Rennes et Sophia (figure 5.2c ou Luxembourg et Sophia (figure 5.2d) ont des profils dont le RTT varie selon les jours de la semaine.

Globalement cependant, les RTTs sont plutôt similaires selon les jours de la semaine durant lesquels ils sont observés (figure 5.2a). En considérant qu'un RTT plus élevé correspond à une plus grande activité, aucun pic global d'activité n'est détecté, contrairement à ce qui peut être observé sur internet [62]. Ce n'est pas anormal, car la bande passante dédiée est suffisamment élevée pour ne pas constituer de goulot d'étranglement. De plus, toujours globalement, la valeur minimale des RTTs est inférieure à 0, donc hors du domaine d'existence des RTTs, indiquant que certaines mesures ont échoué.

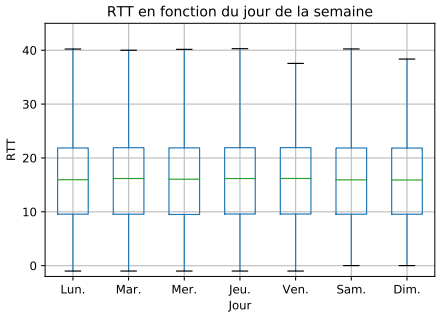
Les mêmes observations sont aussi possibles lorsque la répartition au cours des heures de la journée est observée (figure 5.4a).

Origine \ Dest.	Grenoble	Lille	Luxembourg	Lyon	Nancy	Nantes	Rennes	Sophia
Grenoble	0,02	13,42	17,63	4,3	15,91	16,72	18,18	10,26
Lille	14,12	0,04	12,05	11,53	10,29	24,2	25,53	17,75
Luxembourg	18,47	12,04	0,04	15,79	3,35	28,51	29,84	22,07
Lyon	3,95	10,36	14,51	0,06	12,98	13,91	15,28	7,45
Nancy	16,65	9,66	2,67	14,06	0,02	26,72	28,03	20,35
Nantes	16,65	24,12	28,41	14,03	26,27	0,03	1,94	20,22
Rennes	17,97	25,75	30,16	15,34	27,9	1,97	0,06	21,53
Sophia	10,29	17,72	22,01	7,68	20,08	20,3	21,65	0,03

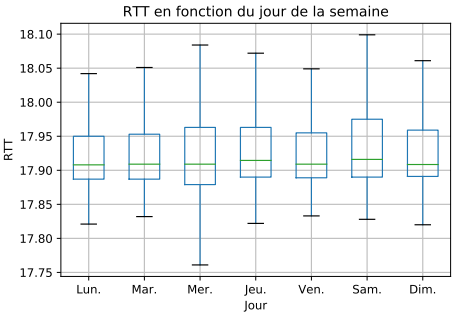
Tableau 5.2 – RTT moyens entre deux nœuds (jeu de données national brut)

Origine \ Dest.	Grenoble	Lille	Luxembourg	Lyon	Nancy	Nantes	Rennes	Sophia
Grenoble	0,01	3,17	3,63	3,88	5,75	0,91	1,39	0,97
Lille	3,72	0,08	1,76	3,71	3,71	4,22	4,36	4,21
Luxembourg	8,1	6,88	0,02	7,93	4,66	8,19	8,16	8,2
Lyon	0,6	2,69	3,07	0,03	4,44	0,42	1,08	0,89
Nancy	4,14	0,81	0,99	4,08	0,01	4,31	4,44	4,27
Nantes	0,67	3,86	4,33	0,64	4,13	0,01	1,42	0,43
Rennes	1,28	4	8,94	0,97	4,49	1,72	0,04	0,9
Sophia	0,63	3,89	4,56	0,6	4,56	0,5	0,8	0,02

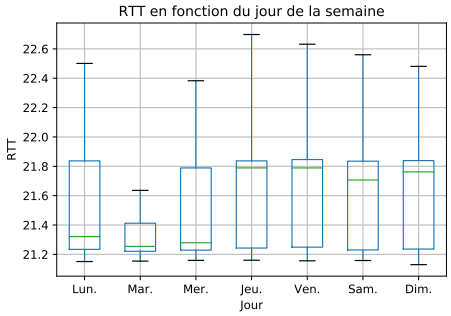
Tableau 5.3 – Écart-types moyens entre deux nœuds (jeu de données national brut)



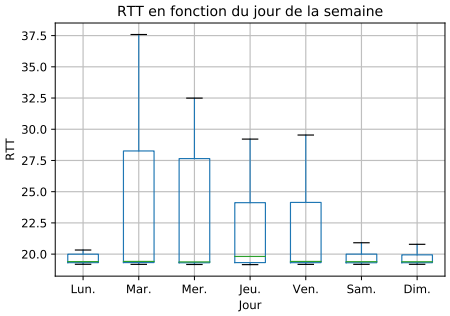
(a) Distribution globale



(b) Distribution entre Rennes et Grenoble



(c) Distribution entre Rennes et Sophia



(d) Distribution entre Luxembourg et Sophia

Figure 5.2 – Distribution du RTT en fonction du jour de la semaine (jeu de données national brut)

Finalement, en observant les nombres de sauts moyens (tableau 5.4), certaines valeurs posent problème. En effet, le nombre de saut sur Grid'5000 est toujours de :

— 1 quand un nœud s'interroge lui-même.

— 3 quand un nœud en interroge un autre : un premier saut jusqu'au routeur local, un second saut jusqu'au routeur distant et un troisième saut jusqu'au nœud interrogé.

C'est aussi un indicateur que certaines mesures ont échoué.

Origine \ Dest.	Grenoble	Lille	Luxembourg	Lyon	Nancy	Nantes	Rennes	Sophia
Grenoble	1	3,13	3,07	3,12	3,96	3,02	3,18	3,02
Lille	3,02	1	3,02	3,02	3,82	3,02	3,08	3,02
Luxembourg	3,02	3,02	1	3,03	3,79	3,03	3,09	3,02
Lyon	3,01	3	3,01	1	4,11	3,01	3,12	3
Nancy	3,02	3,01	3,01	3,03	1	3,03	3,11	3,02
Nantes	3,01	3,01	3,02	3,01	3,82	1	3,07	3,01
Rennes	3,07	3,08	3,1	3,09	3,99	3,06	1	3,08
Sophia	3,01	3,02	3,02	3,01	3,81	3,01	3,08	1

Tableau 5.4 – Nombre de sauts moyens entre deux nœuds (jeu de données national nettoyé)

5.3.2 Nettoyage du jeu de données

Le jeu de données brut présentait différentes anomalies. Pour les résoudre, les nœuds les moins disponibles (Lyon et Nancy) ont été retirés, en tant qu'origine et destination des mesures. De plus les valeurs de RTT impossibles ainsi que les nombres de saut anormaux ont aussi été retirés du jeu de données. Afin de faire fonctionner au mieux les différents algorithmes par la suite, le nombre de mesures a été équilibré, c'est-à-dire que pour une destination donnée, toutes les origines ont le même nombre de mesures. En effet, pour tester les méthodes dans un environnement optimal, l'ensemble des mesures vers une cible doit être disponible.

Le jeu de données nettoyé comporte 172 392 mesures dont la répartition selon l'origine et la destination sondée est détaillée au sein du tableau 5.5.

Origine \ Dest.	Grenoble	Lille	Luxembourg	Nantes	Rennes	Sophia	Total
Grenoble	4557	5141	4712	4778	4768	4776	28732
Lille	4557	5141	4712	4778	4768	4776	28732
Luxembourg	4557	5141	4712	4778	4768	4776	28732
Nantes	4557	5141	4712	4778	4768	4776	28732
Rennes	4557	5141	4712	4778	4768	4776	28732
Sophia	4557	5141	4712	4778	4768	4776	28732
Total	27342	30846	28272	28668	28608	28656	172392

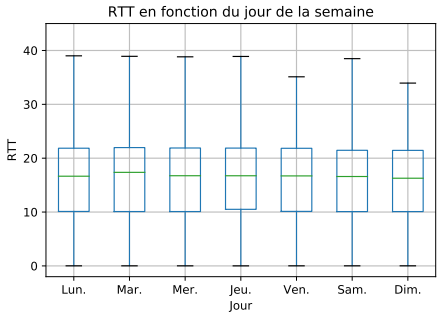
Tableau 5.5 – Nombre de mesures collectées (jeu de données national nettoyé)

Origine \ Dest.	Grenoble	Lille	Luxembourg	Nantes	Rennes	Sophia
Grenoble	0,02	13,54	17,73	16,73	18,1	10,27
Lille	13,51	0,04	11,93	23,54	24,87	17,09
Luxembourg	17,58	11,8	0,04	27,63	28,97	21,16
Nantes	16,63	23,44	27,67	0,03	1,88	20,2
Rennes	17,93	24,79	28,94	1,91	0,06	21,47
Sophia	10,24	17,05	21,23	20,26	21,61	0,03

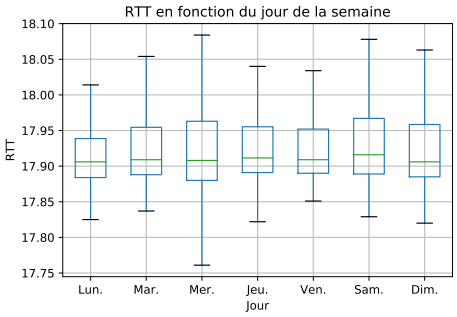
Tableau 5.6 – RTT moyens entre deux nœuds (jeu de données national nettoyé)

Ces modifications du jeu de données ne changent pas la propriété de similarité entre les couples (*origine*; *destination*), mais la mettent en évidence. Il est possible de l'observer au sein du tableau 5.6.

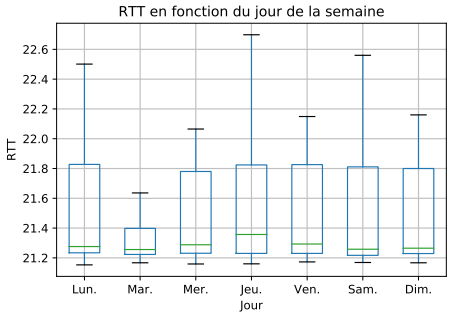
Le jeu de données nettoyé ne permet pas non plus de détecter de pic d'activité selon les jours de la semaine (figure 5.3) ni selon les heures de la journée (figure 5.4b.



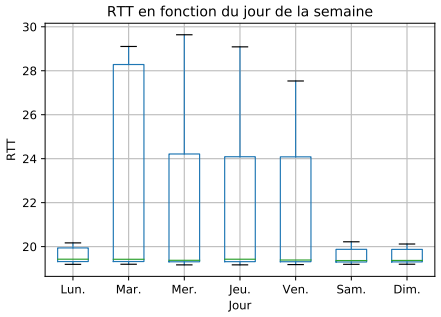
(a) Distribution globale



(b) Distribution entre Rennes et Grenoble



(c) Distribution entre Rennes et Sophia

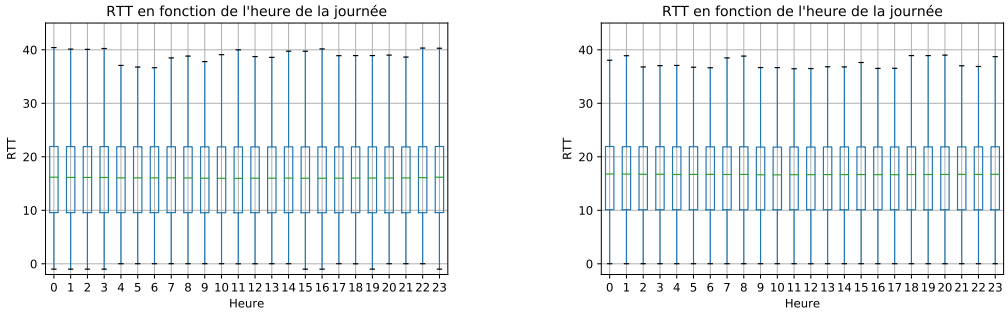


(d) Distribution entre Luxembourg et Sophia

Figure 5.3 – Distribution du RTT en fonction du jour de la semaine (jeu de données national nettoyé)

5.4 Collecte au niveau mondial

Pour établir le premier jeu de données dont les mesures sont réalisées au sein d'un environnement non-contrôlé, nous avons choisi de collecter des mesures au niveau mondial. En effet, en utilisant Amazon AWS, il n'y avait qu'un seul



(a) Jeu de données brut **(b)** Jeu de données nettoyé

Figure 5.4 – Distribution du RTT en fonction de l'heure de la journée (jeu de données national)

site en France. Réaliser une collecte au niveau national, avec un environnement non-contrôlé n'était pas possible avec Amazon, car le seul pays avec plusieurs sites était les États-Unis. Cependant, ce pays est environ 18 fois plus grand que la France et ne comporte que 4 sites Amazon, la comparaison avec l'environnement contrôlé aurait été compliquée. Une autre solution était possible, l'utilisation de PlanetLab [63], qui dispose de plusieurs sites en France, mais l'accès nous y était impossible. Pour accéder à PlanetLab, un site, composé d'au moins deux nœuds, aurait du être ouvert au niveau du laboratoire. Cependant, les conditions d'administration de PlanetLab n'étaient pas compatibles avec celles de l'IRIT.

Tester les différentes méthodes au sein d'un environnement non-contrôlé, c'est-à-dire Internet, permet de répliquer au mieux les conditions réelles d'utilisation des solutions.

5.4.1 Choix et répartition des points de repère

Plateforme AWS-EC2

Les mesures ont été réalisées sur la plateforme AWS à l'aide de machines virtuelles EC2 (VM) [64]. À l'heure actuelle, les services AWS sont disponibles sur 26 sites répartis mondialement mais concentrés en Europe, en Amérique du Nord et en Asie Pacifique [65]. Contrairement à Grid'5000, chaque VM est connecté à Internet, permettant d'interroger des services à l'extérieur du réseau Amazon. Bien sûr, chaque VM propose le même environnement logiciel et des configurations homogènes.

Pour utiliser les services, les VM doivent être instanciées et sont facturées

à la durée d'utilisation, avec un prix variant selon la configuration demandée.

Points de repères sélectionnés

Sur les 26 sites existants nous avons choisi les 15 suivants : Californie du Nord, Canada, Francfort, Irlande, Londres, Mumbai, Ohio, Oregon, Paris, Singapour, Sydney, São Paulo, Séoul, Tokyo et Virginie du Nord. Effectivement, certains sites n'existaient pas lors de la collecte, et d'autres sont réservés pour un usage local, nous avons donc choisi le maximum de sites disponibles. La répartition des points de repère est représentée sur la figure 5.5.

Figure 5.5 – Répartition des nœuds Amazon

5.4.2 Métriques collectées

Similairement au jeu de données national, le RTT et le nombre de sauts a été collecté.

En plus de ces deux métriques, chaque mesure est accompagné d'un horodatage indiquant le moment de la collecte.

5.4.3 Fonctionnement de la collecte

La collecte a duré du 12 juin 2018 au 31 août 2018, soit environ deux mois et demi, avec un rythme d'une mesure toutes les dix minutes. Même en diminuant la fréquence entre deux mesures par rapport au jeu de données précédent, la durée totale permet d'obtenir un jeu de données statistiquement représentatif du réseau.

Un programme similaire à celui utilisé pour le jeu de données national a été mis en place. Ainsi, la liste des nœuds à interroger et l'utilisation de la commande *traceroute* par TCP est conservé. Cependant, il y a une différence, car les VM ne s'interrogent pas entre elles mais vont interroger des sites web dont la position géographique est connue, par exemple des sites d'université. En effet, il n'était pas possible pour les VM d'exécuter des *traceroute* entre elles, bien qu'il ait été possible de le faire par le passé [66–68], le procédé est maintenant bloqué par Amazon. Nous avons appris plus tard qu'utiliser *traceroute* par UDP aurait pu fonctionner [69]. Il était cependant possible d'éxecuter *traceroute* par TCP pour sonder un nœud extérieur.

Ce programme a été lancé sur l'ensemble des 15 VM instanciées chez Amazon.

5.5 Jeu de données mondial

5.5.1 Présentation du jeu de données initial

Le jeu de données mondial comporte 3 643 673 mesures, réalisées depuis 15 VM Amazon EC2 et vers un ensemble de 58 cibles, composées de sites web d'université et de villes, hébergé à l'endroit correspondant. Compte tenu du nombre de cibles, les données présentées dans cette section ne comportent que 9 des 58 cibles.

La répartition selon l'origine et la destination sondée est détaillée au niveau du tableau 5.7. Cette répartition met en évidence le manque de mesures en provenance de Londres, Séoul et de la Virginie du Nord. De plus les nœuds en Irlande et Oregon ont récolté le double des autres nœuds. Ce phénomène s'explique car certains nœuds ont arrêté de mesurer plus tôt que les autres à cause d'un problème de stabilité dans le programme de collecte.

Les RTTs moyens, décrits par le tableau 5.8, varient avec la distance qui sépare le point de repère et les cibles. Plus la distance est élevée, plus le RTT moyen le sera, et inversement. Cependant, interroger certaines cibles, par exemple l'université du Texas, donnait lieu à des valeurs incohérentes. En effet, tous les points de repères mesuraient un RTT moyen trop faible. En considérant que le site était bien hébergé au Texas, il aurait nécessité que la vitesse de traversée de l'information dans le réseau soit plus rapide que celle de la lumière dans la fibre [56]. Le site de l'université était en fait hébergé par Amazon et son contenu dupliqué sur différents points de présence, ce qui faussait nos mesures.

De la même manière qu'avec le jeu de données national, les RTTs mesurés

Origine \ Dest.	Bangkok (u)	Berlin (u)	Brasilia (u)	Istambul (u)	Montpellier (u)	Montpellier (v)	New York (u)	New Delhi (u)	Texas (u)
Cal. du Nord	5617	5616	5617	5617	5617	5617	5617	5617	5617
Canada	1921	1921	1922	1922	1922	1922	1922	1922	1922
Francfort	4624	4624	4625	4625	4625	4625	4624	4625	4625
Irlande	10281	10281	10282	10281	10282	10281	10281	10282	10282
Londres	**210**	**210**	**211**	**210**	**211**	**211**	**210**	**211**	**211**
Mumbai	2357	2356	2357	2357	2357	2357	2357	2357	2357
Ohio	5504	5504	5505	5504	5505	5505	5504	5504	5505
Oregon	11558	11558	11558	11558	11558	11558	11558	11558	11558
Paris	5858	5858	5859	5858	5858	5858	5858	5858	5859
Singapour	4754	4754	4755	4755	4755	4755	4754	4754	4755
Sydney	2979	2979	2980	2980	2980	2980	2979	2979	2980
São Paulo	3997	3997	3997	3997	3997	3997	3997	3997	3997
Séoul	**583**	**583**	**584**	**584**	**584**	**584**	**583**	**583**	**584**
Tokyo	2257	2257	2258	2258	2258	2258	2257	2258	2258
Vir. du Nord	**318**	**318**	**319**	**319**	**319**	**319**	**319**	**319**	**319**

Tableau 5.7 – Nombre de mesures collectées (jeu de données mondial brut)

Origine \ Dest.	Bangkok (u)	Berlin (u)	Brasilia (u)	Istambul (u)	Montpellier (u)	Montpellier (v)	New York (u)	New Delhi (u)	Texas (u)
Cal. du Nord	208,47	169,23	213,47	209,94	160,42	158,81	73,2	275,43	**2,2**
Canada	271,92	127,93	171,57	150,45	102,68	102,57	21,08	222,8	**9,67**
Francfort	211,97	12,89	250,52	47,52	24,22	24,55	92,84	152,94	**1,27**
Irlande	216,07	31,38	209,28	77,28	30,59	30,65	80,03	166,98	**11,24**
Londres	208,12	21,7	230,87	60,04	21,31	21,4	70,87	148,94	**1,55**
Mumbai	102,01	126,38	409,55	170,82	123,61	123,89	245,19	36,66	**3,4**
Ohio	248,17	124,62	181,91	150,43	115,77	113,8	18,4	225,61	**16,58**
Oregon	226,75	184,4	201,08	229,08	168,88	169,18	72,35	294,57	**7,12**
Paris	204,91	20,69	234,63	55,82	14,48	14,84	82,97	152,51	**0,83**
Singapour	32,09	343,18	395,34	289,22	279,44	282,92	246,13	91,79	**2,82**
Sydney	151,43	342,77	341,36	352,32	321,89	322,13	248,11	354,55	**1,67**
São Paulo	392,58	226,24	21,77	284,46	218,2	217,72	129,34	354,2	**5,39**
Séoul	132,79	300,68	352,43	338,38	283,19	283,31	209,4	181,34	**32,36**
Tokyo	103,92	279,84	322,59	317,1	284,05	279,13	180,93	150,4	**5,55**
Vir. du Nord	267,78	104,76	175,99	141,32	94,03	93,94	7,54	223,52	**0,76**

Tableau 5.8 – RTT moyens entre deux nœuds (jeu de données mondial brut)

ne varient pas en fonction d'un pic d'activité horaire ou journalier.

5.5.2 Nettoyage du jeu de données

Afin de nettoyer le jeu de données, les éléments problématiques ont été retirés. D'abord les VM ne réalisant pas assez de mesures et qui ne peuvent donc pas être considérées comme fiable ont été exclues ; ce sont celles situés à Londres, Séoul et en Virginie du Nord. Ensuite, les différentes valeurs de RTT problématiques, telles que celles impliquant de dépasser la vitesse de propagation de la lumière ont été retirées du jeu de données. Éliminer ces valeurs a impliqué de retirer le site de l'université du Texas de la liste des cibles car toutes les mesures dont il était la cible impliquaient cela. De plus, les valeurs de RTT aberrantes (outliers) restantes ont été éliminées en sélectionnant pour chaque couple (*origine*; *destination*) les valeurs se situant à deux écart-types de la moyenne du RTT. Finalement, les données de la fin de la collecte ont été retirées, car les mesures disponibles ne provenaient que de deux nœuds.

Ce nettoyage n'a pas été aussi évident que celui du jeu de données Grid'5000. En visualisant les données, certaines valeurs aberrantes ont pu être mises en évidence assez rapidement. Cependant, compte tenu du nombre important de mesures, tout n'a pas pu être visualisé en même temps, seulement un jeu réduit d'échantillons à partir duquel les observations ont été généralisées. Ainsi certaines mesures problématiques, comme, les mesures vers l'université du Texas, n'ont été détectée qu'après avoir lancé l'évaluation et en essayant d'interpréter les résultats.

Après nettoyage, le jeu de données comporte 1 197 575 éléments. Le tableau 5.9 indique la répartition selon l'origine et la destination pour les cibles sélectionnées. Il n'a pas été équilibré, mais il l'est quasiment. Nous voulions ainsi nous rapprocher des conditions expérimentales réelles, c'est-à-dire qu'une mesure en provenance d'un point de repère peut ne pas être disponible de temps à autres.

Ce nettoyage du jeu de données n'a pas changé ses propriétés statistiques, les moyennes sont similaires à celles qui étaient observables avant le nettoyage, comme indiqué par le tableau 5.10.

Origine \ Dest.	Bangkok (u)	Berlin (u)	Brasilia (u)	Istambul (u)	Montpellier (u)	Montpellier (v)	New York (u)	New Delhi (u)
Cal. du Nord	1895	1917	1918	1882	1901	1916	1773	1918
Canada	1886	1882	1910	1848	1921	1860	1897	1882
Francfort	1890	1697	1915	1829	1901	1901	1918	1920
Irlande	1874	1738	1838	1896	1903	1910	1919	1790
Mumbai	1888	1841	1872	1856	1907	1902	1861	1827
Ohio	1879	1875	1915	1810	1856	1863	1876	1875
Oregon	1891	1881	1881	1854	1853	1874	1920	1903
Paris	1913	1695	1908	1821	1908	1829	1869	1887
Singapour	1910	1915	1919	1859	1895	1891	1737	1728
Sydney	1897	1919	1917	1841	1920	1921	1896	1896
São Paulo	1914	1907	1918	1843	1843	1843	1909	1869
Tokyo	1885	1918	1868	1864	1823	1920	1898	1793

Tableau 5.9 – Nombre de mesures collectées (jeu de données mondial nettoyé)

Origine \ Dest.	Bangkok (u)	Berlin (u)	Brasilia (u)	Istambul (u)	Montpellier (u)	Montpellier (v)	New York (u)	New Delhi (u)
Cal. du Nord	207,02	170,18	213,55	209,31	159,91	160,38	74,87	276,88
Canada	271,83	127,6	169,41	150,83	102,06	102,35	21,0	221,45
Francfort	213,02	11,8	251,52	46,54	24,1	24,45	94,9	149,98
Irlande	216,02	31,0	239,35	75,68	30,2	30,52	78,95	159,99
Mumbai	97,75	126,15	409,39	170,13	123,51	123,78	245,08	25,26
Ohio	249,14	127,08	180,68	147,77	114,28	114,62	17,63	228,57
Oregon	226,96	183,74	230,3	227,31	170,66	171,04	73,88	294,3
Paris	205,49	19,57	235,94	54,62	14,37	14,75	84,42	146,23
Singapour	31,35	341,58	391,19	290,66	278,12	282,61	244,57	79,84
Sydney	150,32	339,89	339,18	349,41	318,78	319,09	247,83	362,43
São Paulo	392,06	227,86	22,01	283,05	215,42	215,7	128,85	366,46
Tokyo	103,8	276,56	321,77	316,29	281,75	277,02	180,82	151,9

Tableau 5.10 – RTT moyens entre deux nœuds (jeu de données mondial nettoyé)

Chapitre 6

Analyse des performances des algorithmes

Sommaire

6.1 Introduction

Les jeux de données collectés et présentés dans le chapitre 5 servent à réaliser l'analyse des différentes méthodes de localisation des données. Ces jeux de données étant déjà collectés pour l'étape d'apprentissage et de vérification ainsi que nettoyés des éléments non fiables, ce chapitre présente leur utilisation dans l'analyse de performance. Pour cela, les jeux de données sont divisés en deux parties : une partie utile à l'apprentissage et une partie utilisée pour la vérification. Différents scénarios de division ont été testés. Lors de la construction de la partie « vérification », les mesures sont regroupées en un jeu de mesure, de telle sorte à ce que pour une cible donnée, le jeu de mesure comporte une mesure réalisée par tout point de repère, autre que celui correspondant à la cible s'il existe, vers cette cible. Par exemple, un jeu de mesure Grid'5000 vers Lille, comprendra une mesure depuis Grenoble vers Lille, une autre depuis Sophia vers Lille, et ainsi de suite pour l'ensemble des points de repère, à l'exception d'une mesure depuis Lille vers Lille.

Par rapport aux méthodes présentées dans le chapitre 3, trois techniques ont été retenues et n'utilisent que les RTTs collectés. Ces techniques sont la bestline, la régression linéaire et la régression polynomiale. Différents degrés de régression polynomiale ont été testés, à savoir des régressions de degré 3, 4 et 5. Ces méthodes ont été retenues, car elles fonctionnent de la même manière, la seule différence étant l'algorithme utilisé pour construire la fonction d'estimation.

Pour chaque méthode, les fonctions d'estimation sont définies avec la partie du jeu de données servant à l'apprentissage. Pour cela, les données sont regroupées selon chaque point de repère d'origine et pour chaque groupe, la fonction d'estimation est calculée, selon l'algorithme défini, avec les valeurs mesurées par le point de repère en question.

Ensuite, un consensus pour chaque ensemble de mesures de la partie vérification est calculé. Pour cela, pour chaque mesure du jeu de mesure, la fonction d'estimation du point de repère correspondant est appliqué, indiquant la distance estimée entre la cible et le point de repère ayant réalisé la mesure. Avec ces distances, un cercle, centré sur le point de repère et dont rayon est la distance estimée, peut être construit pour chaque point de repère. Le consensus représente la zone d'intersection entre plusieurs de ces cercles, c'est-à-dire le résultat d'une multilatération. La sélection du consensus dépend du jeu de données utilisé, mais dans tous les cas, un consensus est représenté par une figure géométrique.

Une fois la multilatération réalisée, pour chaque consensus établi, les scores peuvent être calculés. Chaque score dépend du consensus mais aussi de la zone

cible dans laquelle le stockage des données est accepté. Ainsi, chaque score a été calculé pour différentes zones cibles. Nous avons choisi comme zones cibles, des cercles centrés sur la cible réelle et dont le rayon varie entre 1 km et 20 000 km.

Afin de visualiser les résultats d'une vérification, un outil affichant les différentes figures géométriques construites au dessus d'une carte du monde a été développé.

6.2 Conditions d'évaluation au niveau national

L'évaluation au niveau national utilise le jeu de données collecté sur Grid'5000.

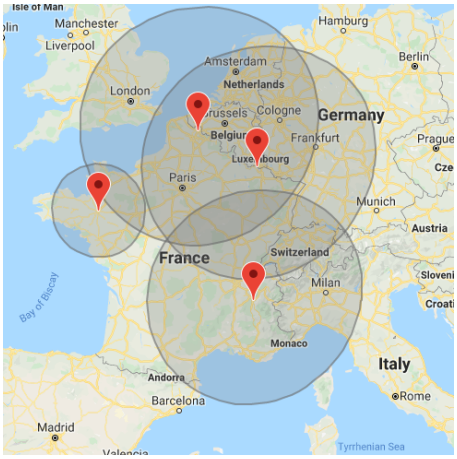
6.2.1 Choix du consensus

Lors de cette évaluation le consensus considéré est le consensus maximum. C'est-à-dire celui regroupant le plus de points de repère. L'environnement au sein duquel les mesures ont été récoltées était contrôlé, le consensus maximum est donc le plus souvent total, celui composé des 5 points de repère. Dans de rares cas, le consensus n'est pas total et est composé uniquement de 4 points de repère. Bien qu'il soit possible, dans les cas où le consensus n'est pas total, que plusieurs solutions soient en concurrence, dans la pratique aucun de ces cas n'a été rencontré. Cette solution permet de ne pas exclure les cas où un point de repère vient fausser le résultat alors que les autres conviennent d'une solution.

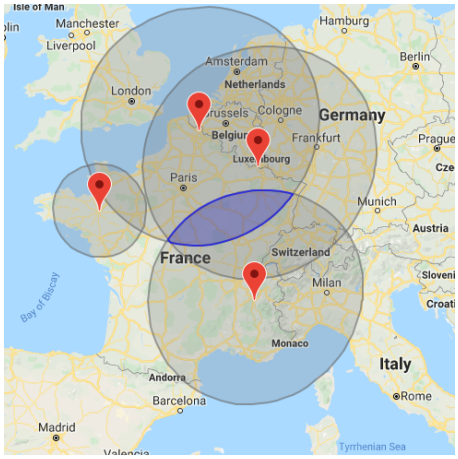
Par exemple, la figure 6.1 illustre cette situation. Quatre points de repère sont disponibles et ont chacun estimé la position des données à une certaines distance d'eux-mêmes, ce qui est représenté par les cercles gris. Sur la figure 6.1a, les 4 cercles ne forment pas une intersection commune, donc le consensus maximum est choisi, c'est à dire la zone bleue représentée sur la figure 6.1b.

6.2.2 Sélection du degré de régression polynomiale

Différents degrés de régression polynomiale ont été utilisés et testés. Cependant seule la régression polynomiale de degré 3 a été conservée. Il y a deux raisons à cela :

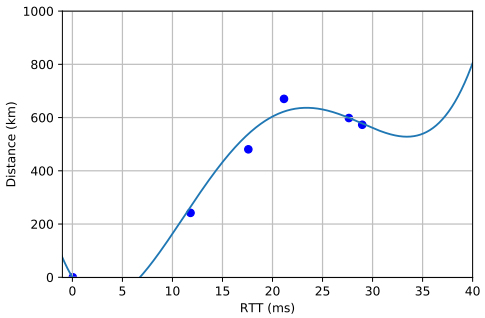


(a) Distance estimée par chacun des points de repère

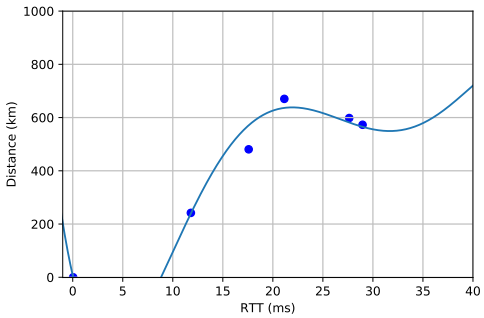


(b) Consensus maximum pour les points de repère

Figure 6.1 – Exemple de consensus maximum



(a) Fonction polynomiale de degré 4



(b) Fonction polynomiale de degré 5

Figure 6.2 – Fonctions polynomiales estimées sur le jeu de données Grid'5000

— Les régressions polynomiales de degré 4 et 5 ne semblaient pas être de bonnes candidates. Dans les deux cas, bien que la régression corresponde aux données, certaines fonctions prédisent des distances négatives pour des RTTs dans l'intervalle $]0, 7]$, ce qui est impossible, comme il peut être observé sur la figure 6.2.

— Les résultats des scores obtenus en utilisant ces fonctions ne différent que légèrement de ceux obtenus par la régression polynomiale de degré 3 (figure 6.3), qui ne présente pas d'anomalies de prédictions.

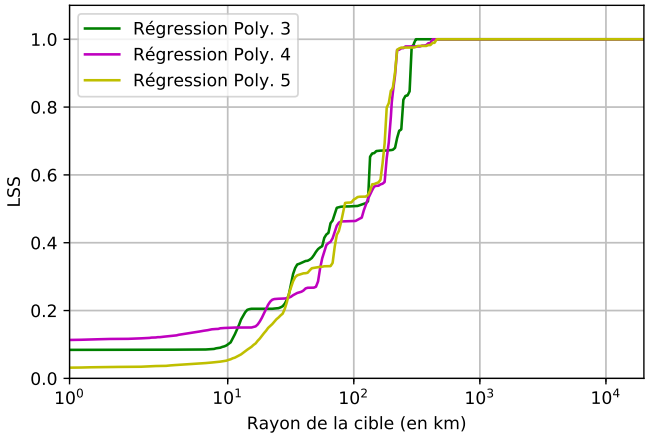


Figure 6.3 – LSS obtenus par les différentes fonctions polynomiales

Sachant que les degrés 4 et 5 ne donnaient pas de résultats intéressants, les degrés supérieurs n'ont pas été testés.

Dans la suite, le terme de régression polynomiale indiquera la régression de degré 3, sauf mention contraire.

6.2.3 Scénarios de division des mesures

Trois scénarios de divisions pour l'apprentissage et la vérification ont été choisis avec l'hypothèse que plus la partie utilisée par l'entrainement est grande, meilleures en seront les prédictions et les scores. La sélection des mesures pour la division est faite aléatoirement. Les scénarios testés sont :

— 0.8/0.2 avec 80% des mesures pour l'apprentissage et 20% pour la vérification et l'évaluation.

— 0.5/0.5 avec 50% des mesures pour l'apprentissage et 50% pour la vérification et l'évaluation.

— 0.2/0.8 avec 20% des mesures pour l'apprentissage et 80% pour la vérification et l'évaluation.

Ces scénarios apportent une vue d'ensemble des différents ratios de division possibles, bien qu'ils ne soient pas exhaustifs.

En testant ces différents scénarios pour une taille de cible donnée, les résultats obtenus sont similaires, c'est-à-dire que les scores ne changent pas quel que soit le ratio de répartition choisi, il ne change pas les scores. Ainsi, nous pouvons choisir n'importe laquelle de ces répartitions sans incidence sur les résultats. Nous avons donc choisi la répartition 0.8/0.2 car l'apprentissage n'est pas coûteux en temps à la différence de la vérification qui nécessite la recherche du consensus et de l'évaluation qui nécessite des calculs d'intersections de surface.

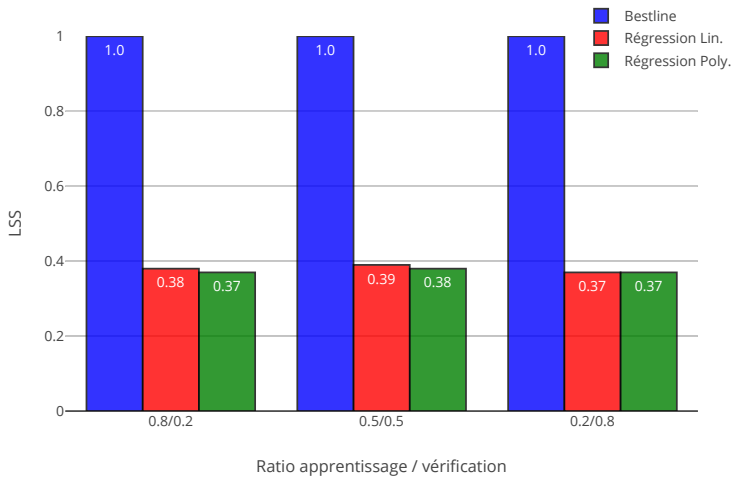


Figure 6.4 – LSS selon différents scénarios de division (rayon cible 50 km)

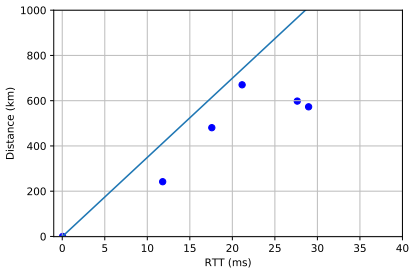
6.3 Présentation des résultats au niveau national

6.3.1 Fonctions d'estimation

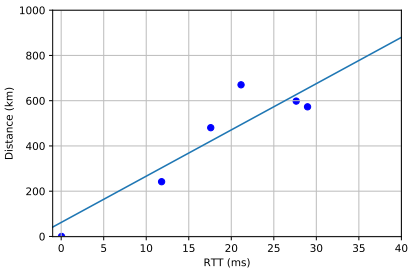
Les fonctions d'estimation obtenues suivent la distribution des données, et respectent les propriétés des fonctions. Certaines de ces fonctions, celles associées au point de repère du Luxembourg, sont tracées sur la figure 6.5. Pour ne pas surcharger les graphiques, les RTTs affichés sont les RTTs moyens par distance. En effet, le nombre de points de repère étant limité, il n'y a que 6 distances différentes, mais chacun présente un nombre important de RTT.

— La fonction bestline (figure 6.5a) surestime les distances : quel que soit le RTT, la distance estimée sera toujours supérieure ou égale à la distance réelle entre les deux points de mesure.

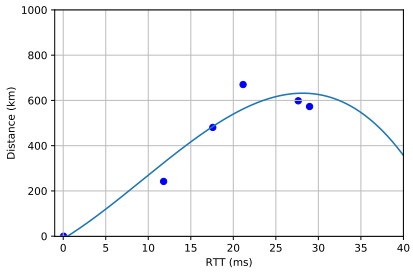
— Les fonctions par régression (figure 6.5b et figure 6.5c) donnent une vision « moyenne » de la distance et la surestiment ou sous-estiment en fonction de la forme des données.



(a) Bestline



(b) Régression linéaire



(c) Régression polynomiale de degré 3

Figure 6.5 – Fonctions estimées sur le jeu de données Grid'5000

6.3.2 LSS

Le premier score est le LSS, représentant le succès de la localisation, c'est-à-dire s'il existe une intersection entre la prédiction de la localisation et la localisation réelle. Il représente aussi la probabilité estimée d'observer une telle intersection. Il permet d'évaluer le succès de la méthode mais n'atteste pas de la qualité du résultat.

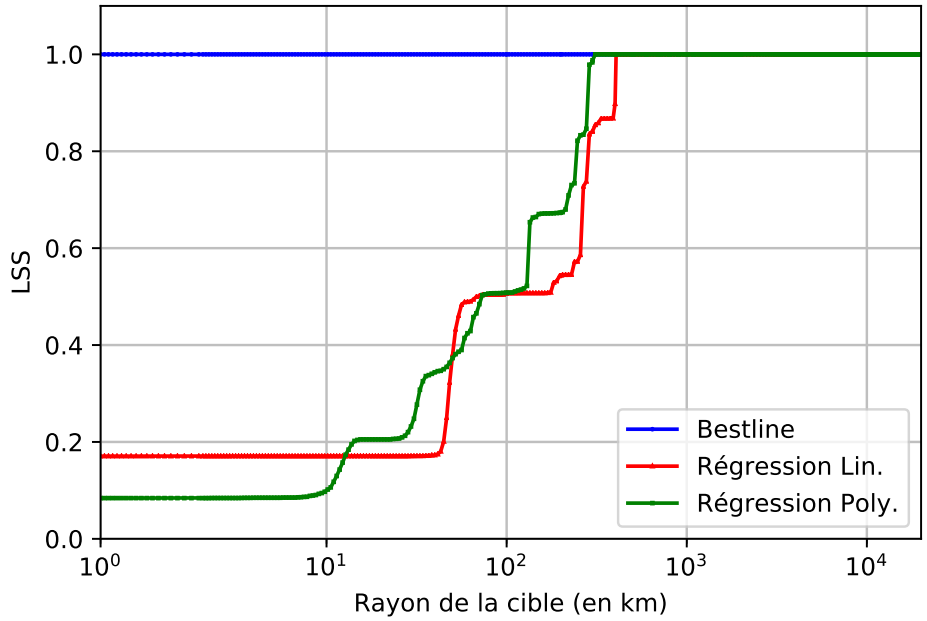


Figure 6.6 – LSS sur le jeu de données Grid'5000

Le LSS de chaque méthode en fonction du rayon du cercle autour de la cible est donné par la figure 6.6. Plusieurs points sont à remarquer :

— Les LSS des différentes méthodes finissent tous par converger à 1. Il y a toujours au moins un consensus, et les consensus et la zone cible sont des figures géométriques évoluant à la surface d'une sphère (la planète Terre). Il est donc toujours possible de trouver une taille de la zone cible finie à partir de laquelle tous les consensus ont une intersection avec la zone cible. De plus, le LSS en fonction de la taille de la zone cible est croissant. En effet, les zones prédites ne changent pas, mais la taille de la zone cible croit, elle finira donc par intersecter les zones prédites.

— La bestline donne toujours un LSS de 1, quel que soit la taille de la zone

cible. Ce résultat n'est pas surprenant, avec la bestline les distances sont surestimées, et si le jeu de données utilisé pour l'apprentissage est représentatif du réseau, les données utilisées pour la vérification donneront toujours un succès en terme de LSS.

— Les deux autres méthodes n'ont pas d'aussi bons résultats, il faut une cible avec un rayon d'environ 70 km pour obtenir un score 0.5. Pour avoir un score de 1, le rayon doit être de 300 km pour la régression polynomiale et 400 km pour la régression linéaire.

6.3.3 CRS

Le second score est le CRS, représentant le ratio de la zone prédite s'intersectant avec la zone cible, lorsque celui-ci existe. Il permet d'évaluer le succès du résultat.

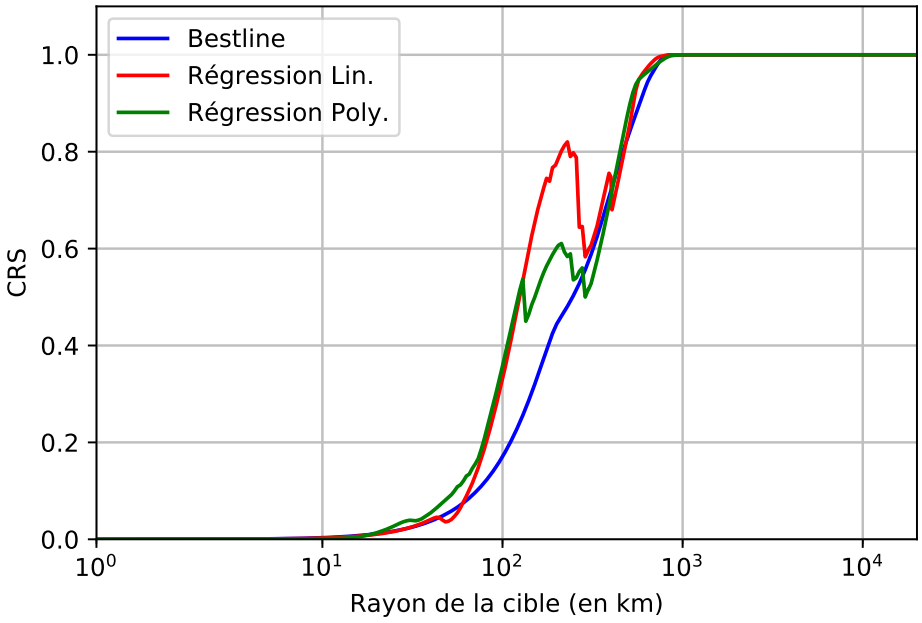


Figure 6.7 – CRS sur le jeu de données Grid'5000

Le CRS de chaque méthode, en fonction du rayon du cercle autour de la cible, est donné par la figure 6.7. Plusieurs points sont à remarquer :

— Comme pour le LSS et pour les mêmes raisons, le CRS des différentes méthodes finissent tous par converger vers 1. La convergence s'effectue

environ au même moment, lorsque le rayon de la cible est d'environ 1000 km.

— Le CRS donné par les méthodes par régression est supérieur à celui donné par la bestline. Bien qu'il soit le plus souvent croissant, il peut décroitre quand le LSS augmente. Aux alentours d'un rayon de la cible de 230 km jusqu'à 300 km, le CRS décroit pour les méthodes par régression. Pour la régression linéaire cela correspond à une baisse du CRS de 0.82 à 0.6, pendant que le LSS augmente de 0.55 à 0.84. En effet, lorsque le LSS augmente, de nouveaux consensus sont disponibles pour le calcul du LSS et ces nouveaux consensus peuvent avoir un CRS « local » faible, diminuant ainsi la moyenne.

— Le CRS donné par la bestline est croissant. En effet, comme le LSS vaut toujours 1, toutes les prédictions sont incluses dans le calcul du CRS, qui ne peut qu'augmenter pour une prédiction donnée quand le rayon de la cible augmente.

6.4 Conditions d'évaluation au niveau mondial

L'évaluation au niveau mondial utilise le jeu de données collecté sur Amazon.

6.4.1 Division apprentissage/vérification et degré de régression polynomiale

Des tests préliminaires avec le jeu de données Amazon ainsi que les conclusions issues de l'expérimentation précédente sur Grid'5000 ont permis d'éliminer les polynômes de degré supérieur à 3 et de conserver uniquement une régression linéaire de degré 3. De plus, toujours dans le but d'accélérer le calcul des scores, un ratio 0.9/0.1 (90% des mesures pour l'apprentissage et 10% pour la vérification et l'évaluation.) a été choisi. En effet, le ratio de divisions des mesures pour l'apprentissage et la vérification n'a pas d'incidence sur le calcul des scores si chaque partie du jeu divisé présente les mêmes caractéristiques. D'après le tableau 6.1, cette situation est respectée.

Origine	Moyenne du RTT		Écart-type du RTT	
	Apprentissage	Vérification	Apprentissage	Vérification
Tokyo	236,03	235,97	74,1	74,09
Mumbai	155,77	155,57	73,25	72,12
Singapour	257,46	257,46	85,56	85,38
Sydney	290,51	290,76	62,54	63,04
Canada	113,91	113,81	57,15	56,23
Francfort	66,66	66,31	77,57	76,88
Irlande	64,86	64,59	65,91	64,78
Paris	61,5	61,47	76,14	76,1
São Paulo	220,79	220,55	64,81	63,97
Ohio	119,8	119,78	57,76	58,26
Caroline du Nord	152,27	152,05	63,92	62,51
Oregon	162,51	162,78	71,08	72,99

Tableau 6.1 – Moyenne et écart-type du RTT entre apprentissage et vérification par point de repère (ratio 0.9/0.1)

6.4.2 Choix du consensus

Avant de compléter l'évaluation du jeu de données Amazon et car celle sur Grid'5000 indiquait que c'était possible, le consensus choisi était le consensus total. Cependant, au sein de ce jeu de données, beaucoup de jeux de mesures ne permettent pas un consensus total. De plus, le nombre de mesures en provenance de chaque point de repère n'est pas équilibré, donc un consensus maximum aurait mélangé des consensus à 3 points de repère et d'autres à 12. Nous avons donc choisi de considérer l'ensemble des consensus existants. C'est-à-dire, pour chaque jeu de mesure, une fois que chaque point repère a été associé à un cercle, toutes les combinaisons de cercles sont testées. Par exemple s'il y a n mesures, donc n cercles, le nombre total de combinaisons à tester est :

$$\sum_{i=1}^{n} \binom{i}{n} = \sum_{i=0}^{n} \binom{i}{n} - \binom{0}{n} = 2^n - 1$$

Comme nous avons 12 points de repère, tester un jeu de mesure, pour une méthode donnée et un rayon de cible donnée, revient au pire cas à calculer $2^{12} - 1 = 4095$ consensus puis le score de chaque consensus. Bien sûr, certaines combinaisons ne sont pas « valides », et donnent une intersection nulle, mais il faut quand même tester la plupart des cas avant de pouvoir en éliminer.

Si l'on reprend l'exemple utilisé précédemment, pour le consensus maximum, avec 4 points de repère, on aurait ici $2^4 - 1 = 15$ consensus de 1 à

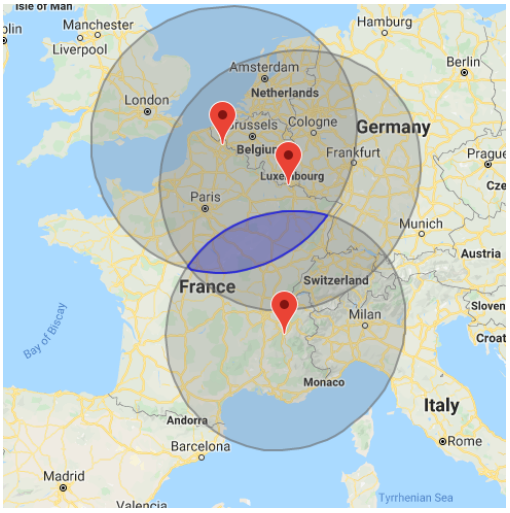
4 points de repère à explorer. Pour 4 points de repère, il y a théoriquement $\binom{4}{4} = 1$ consensus, mais les 4 cercles ayant pour origine les points de repère ne forment pas d'intersection commune. Avec 3 points de repère, il y a $\binom{3}{4} = 4$ consensus théoriques, mais un seul existe, le consensus affiché sur la figure 6.8a. Ensuite, en considérant les consensus à 2 points de repère, leur nombre théorique est $\binom{2}{4} = 6$, mais sur ces 6 consensus, seulement 3 sont possibles, ceux illustrés par les figures 6.8b, 6.8c et 6.8d. Finalement, avec 1 point de repère, il ya $\binom{1}{4} = 4$ consensus, qui sont chacun représenté par le cercle originant du point de repère. Contrairement au nombre théorique de 15 consensus dénombrés, seulement 8 existent.

6.4.3 Estimation des intersections par une méthode de type Monte-Carlo

Rechercher un consensus et calculer un score revient essentiellement à réaliser des intersections de figures géométriques. Cependant, dans notre implémentation, le calcul des intersections est un calcul coûteux. En effet, la seule bibliothèque capable de manipuler ce genre de figures, à notre connaissance, est Shapely [70,71]. Si le temps pour réaliser une intersection avec cette bibliothèque est généralement raisonable, le nombre d'intersection à calculer rend le temps total d'éxecution trop élevé. De plus, plusieurs des consensus théoriques n'existent pas, mais le seul moyen de le savoir de manière « classique » est de tenter de les calculer. Afin d'obtenir des résultats dans un temps raisonnable, en plus de paralléliser le code, il était nécessaire de trouver un moyen pour accélérer le calcul des intersections.

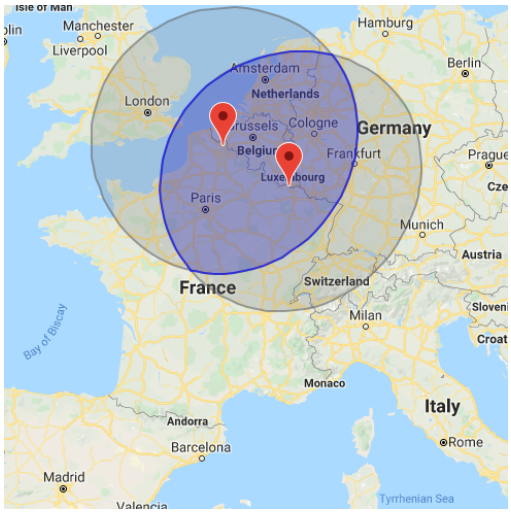
Pour cela, une méthode de type Monte-Carlo a été employée. Il s'agit de considérer le cercle construit autour de la cible et de tirer des points aléatoirement à l'intérieur de celui-ci. Les points sont choisis uniquement à l'intérieur du cercle car les points à l'extérieur du cercle ne sont pas intéressants pour le calcul des scores.

Le nombre de points à tirer dépend de l'aire du cercle autour de la cible, plus le rayon de ce cercle est grand : plus il faut de points pour approximer les intersections. Nous avons donc décidé d'une pseudo-densité proportionnelle au diamètre du cercle. En effet, la densité réelle est proportionnelle au carré du rayon du cercle. Cependant, les propriétés des consensus permettent de réduire la densité nécessaire pour estimer les intersections lorsque le rayon du cercle augmente. Réduire la densité permet de limiter le nombre de point à tirer, pour ne pas se retrouver dans une situation qui est elle aussi coûteuse lorsque le rayon du cercle associé à la zone cible augmente.

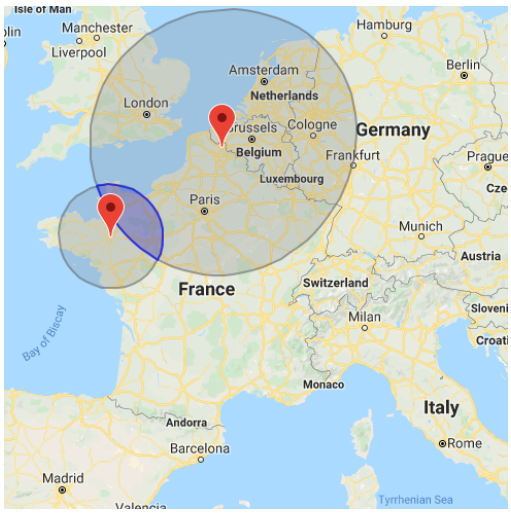


(a) Consensus entre Grenoble, Lille et Luxembourg

(b) Consensus entre Grenoble et Luxembourg



(c) Consensus entre Lille et Luxembourg



(d) Consensus entre Rennes et Lille

Figure 6.8 – Exemple de différents consensus

Pour chaque point, il est facile de déterminer s'il appartient individuellement à chaque cercle construit autour des points de repère. Ainsi, pour chaque point, il est possible de construire le consensus dans lequel il se trouve. En utilisant assez de points, il est possible d'estimer correctement l'ensemble des consensus existants. De plus, quelques propriétés garantissent un peu plus la précision de l'estimation :

— Si un point fait partie d'un consensus formé par k cercles, alors il fait aussi partie de tous les consensus de $k-1$ cercles. En effet, si un point fait partie d'un consensus de k cercles, il fait partie des k cercles donc de chaque combinaison de $k-1$ de ces cercles.

— Si un consensus existe pour un rayon de cible r_i alors ce consensus existe aussi pour tout rayon de cible $r_j \geq r_i$. En effet, si une intersection entre un consensus avec un cercle de rayon donné existe, alors l'intersection entre ce consensus et un cercle de même centre mais de rayon supérieur existe aussi.

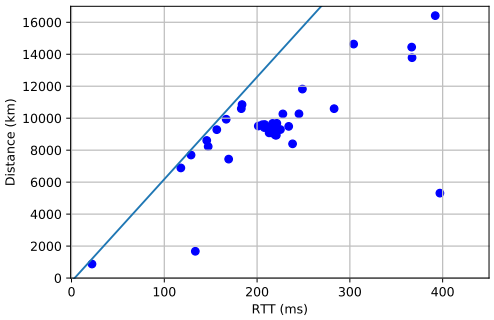
Cependant, il reste le problème des consensus non-existants qui ne sont pas détectés avec cette méthode et interférent avec l'estimation du LSS. En effet, pour calculer le LSS pour une taille de cible donnée et un nombre de points de repère dans le consensus donné (i) avec un nombre de points de repère connu (n), il suffit de calculer le ratio :

$$\frac{\text{nombre de consensus détectés}}{\text{nombre théorique de consensus}} = \frac{\text{nombre de consensus détectés}}{\binom{i}{n}}$$

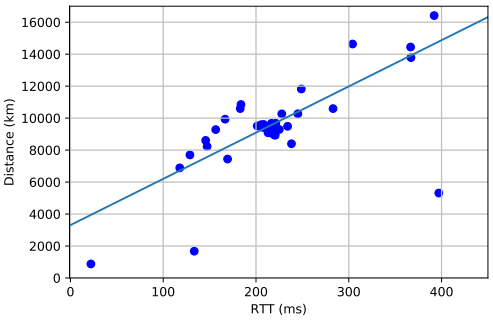
Mais il est possible qu'un consensus n'existe pas, car les cercles qui devraient le composer ne forment pas d'intersection et soit comptabilisé dans le calcul du score. Ces consensus ne sont pas différenciables des consensus existants mais n'ayant pas d'intersection avec la cible. Nous faisons donc l'hypothèse que ces consensus sont en nombre assez faible pour ne pas interférer avec le calcul du score.

6.5 Présentation des résultats au niveau mondial

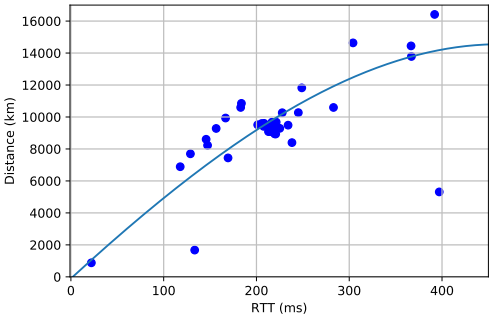
6.5.1 Fonctions d'estimation



(a) Fonction obtenue par bestline



(b) Fonction obtenue par régression linéaire)



(c) Fonction obtenue par régression polynomiale de degré 3

Figure 6.9 – Fonctions estimées sur le jeu de données Grid'5000

Comme lors de l'évaluation des données de Grid'5000, les fonctions d'estimation obtenues suivent la distribution des données, et respectent les propriétés des fonctions. Certaines de ces fonctions, celles associées au point de repère de São Paolo, sont tracées sur la figure 6.9. Pour ne pas surcharger les graphiques, les RTTs affichés sont les RTTs moyens par distance car chaque distance présente un nombre important de RTT.

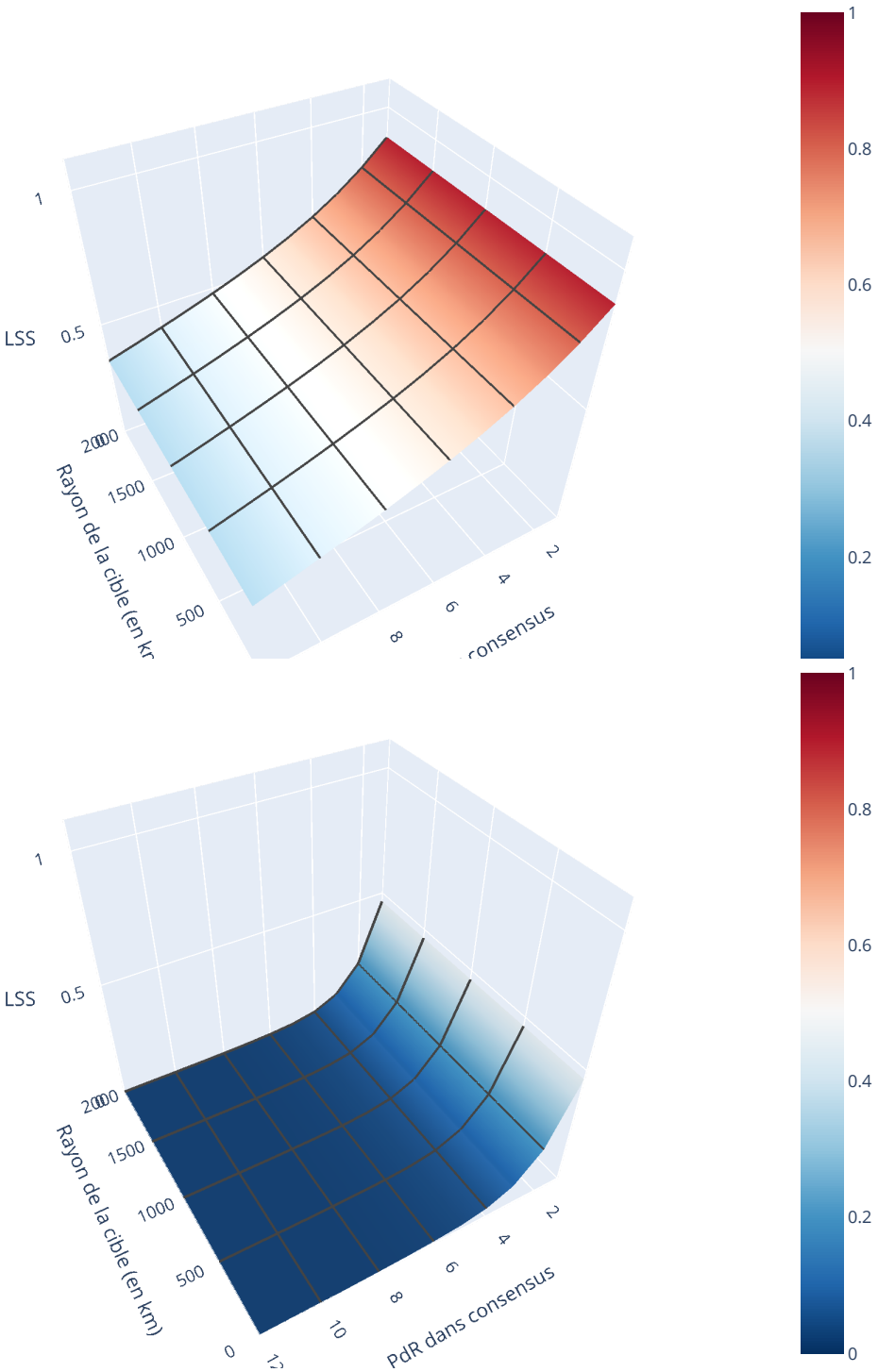
6.5.2 LSS

Le LSS estimé par Monte-Carlo en fonction du rayon du cercle autour de la cible et du nombre de points de repère dans le consensus est donné pour chaque méthode par la figure 6.10. Plusieurs points sont à remarquer :

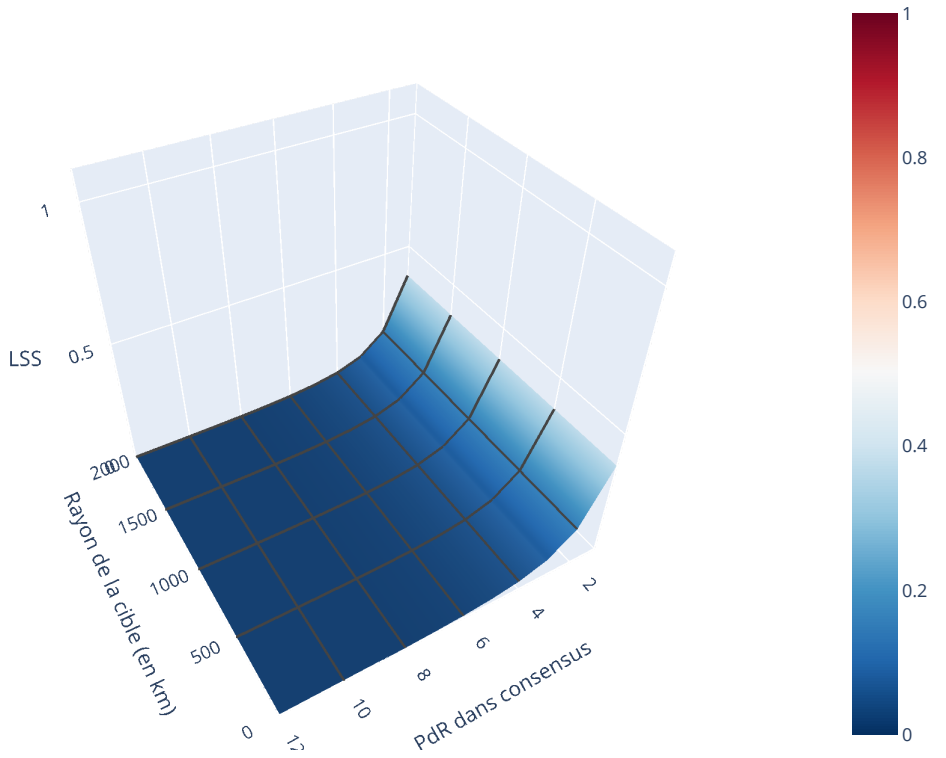
— Le LSS de chaque méthode augmente lorsque le nombre de points de repère dans le consensus diminue. Il y a deux raisons à cela, d'abord, il est plus facile d'avoir un consensus avec peu de cercles qu'un consensus avec beaucoup de cercles, il y a donc plus de consensus valides avec peu de cercles qu'avec beaucoup de cercle par rapport au nombre théorique. Ensuite, les consensus avec peu de cercles définissent une zone généralement plus étendue que ceux avec plus de cercles, renforçant les chances d'intersection et donc d'augmenter le LSS.

— Toujours lorsque le nombre de points de repère dans le consensus diminue, le comportement de la méthode bestline (figure 6.10a) est différent par rapport aux deux autres (figures 6.10b et 6.10c) qui sont identiques. Au niveau de la bestline, le LSS évolue linéairement de 0.34 à 0.89 lorsque le nombre de points de repère diminue. Alors que, pour les deux autres techniques, l'évolution est quasiment nulle de 12 à 4 points de repère, avec un score évoluant de 0 à 0.03. Puis l'évolution devient exponentielle, pour atteindre 0.37 pour la régression polynomiale et 0.46 pour la régression linéaire.

— Sur le domaine observé, changer le rayon de la cible ne modifie pas le LSS mais cela est peut être dû au domaine limité. En effet, au-delà de 2000 km de rayon, la zone cible devient trop importante pour être réaliste et les résultats précédents sur les données issues de Grid'5000 indiquaient des scores limités à 1 après 1000 km de rayon, nous avons donc choisi de ne pas étendre le rayon au delà. Une autre raison est que le score maximum est déjà atteint à cause des consensus non-existants comptabilisés dans le calcul du LSS.



(b) LSS pour régression linéaire



(c) LSS pour régression polynomiale

Figure 6.10 – LSS en fonction du nombre du rayon de la cible et du nombre de points de repère dans le consensus sur le jeu de données Amazon

6.5.3 CRS

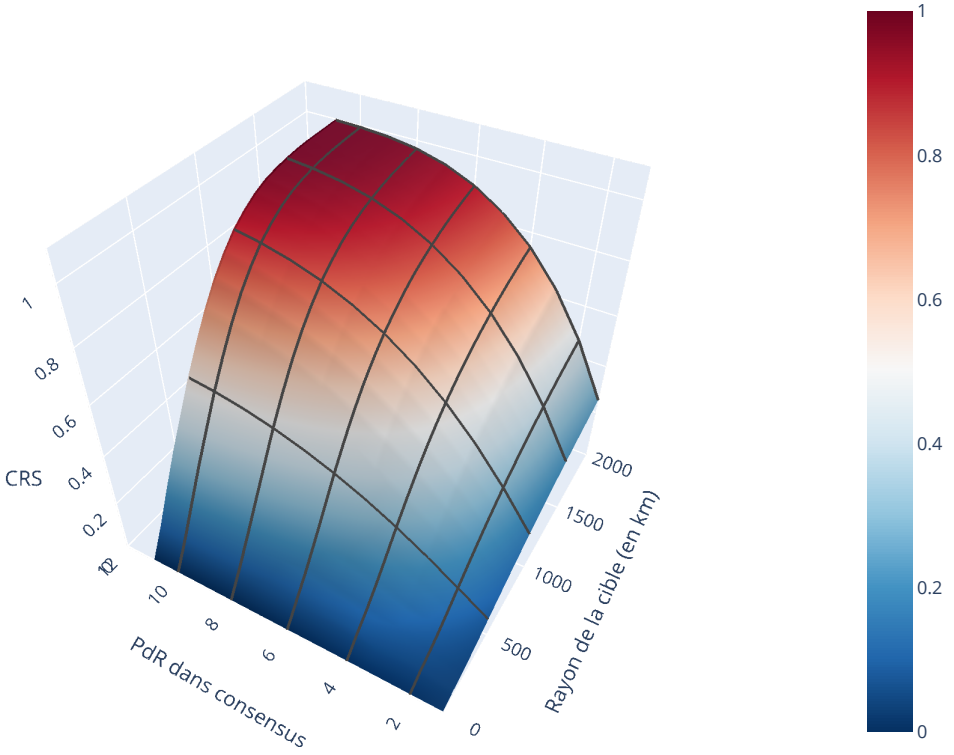
Le CRS estimé par Monte-Carlo en fonction du rayon du cercle autour de la cible et du nombre de points de repère dans le consensus est donné pour chaque méthode par la figure 6.10. Plusieurs points sont à remarquer :

— Le CRS de chaque méthode croit avec le rayon de la cible. Les consensus existant avec un rayon faible ont un CRS « local » plus important, cela signifie donc que les nouveaux consensus, qui existent, car le LSS augmente selon cet axe aussi, sont soit assez peu nombreux pour inverser le gain qu'apporte l'augmentation du rayon, ou bien leur CRS « local » est déjà important.

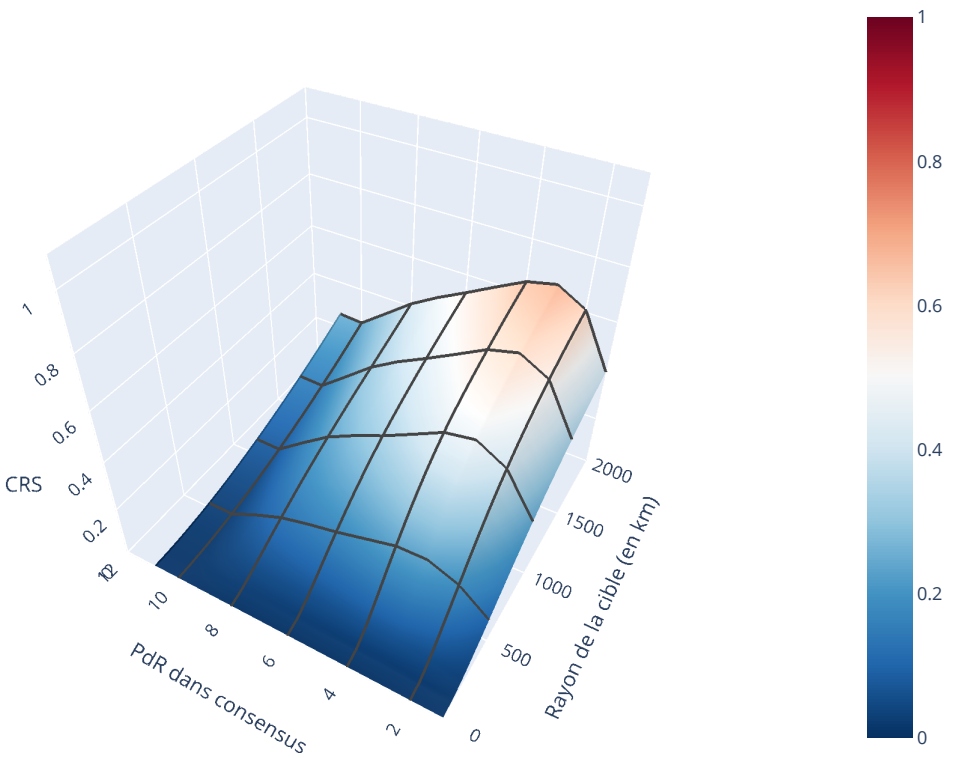
— Pour la méthode bestline, le CRS croit aussi avec le nombre de points de repère dans le consensus, pour atteindre 1 quand le consensus est maximal et la distance de la cible aussi. En effet, la méthode bestline

surestime les distance, et augmenter le nombre de points de repère dans le consensus diminue la taille de ces consensus, donc, pour un rayon de cible de donné, le score augmente car la diminution de la taille s'effectue sur la partie du consensus hors du cercle représentant la cible.

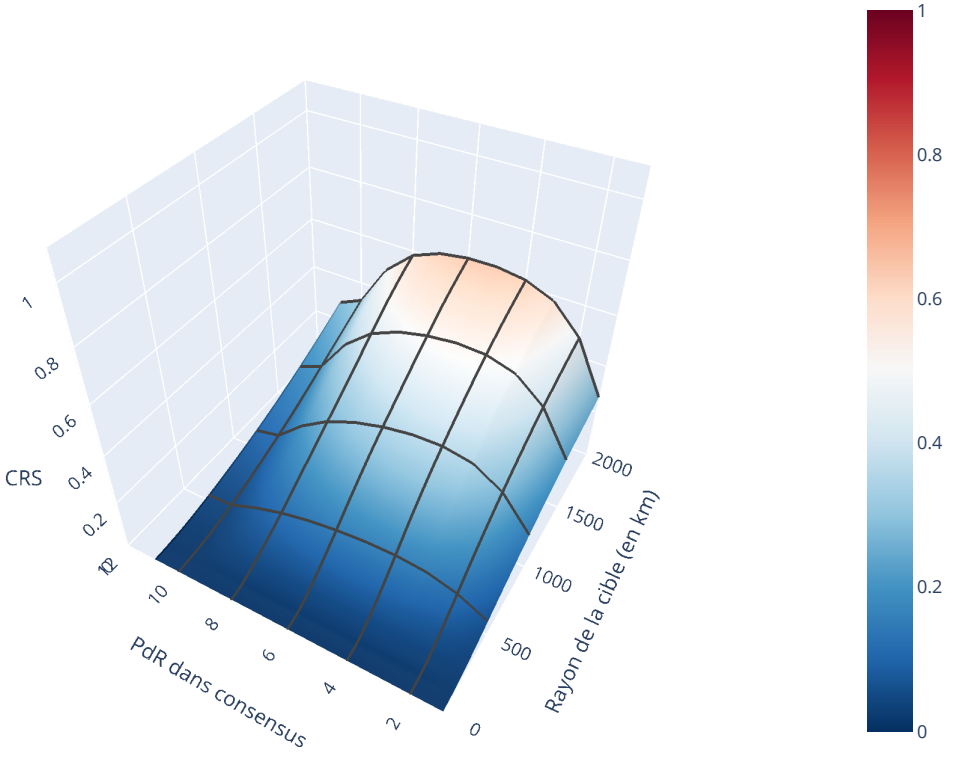
— Pour les méthodes par régression, le CRS croit pour certains consensus, ceux situés plutôt vers le milieu de l'axe et diminue aux valeurs extrêmes de consensus. La plupart des consensus dont le CRS « local » est calculé sont ceux dont le LSS « local » vaut 1. Cependant, d'après le LSS, très peu voire aucun consensus avec plus de 4 points de repère sont impliqués dans ce calcul. C'est pour cela que le CRS augmente d'abord, quand un nombre suffisant de consensus est impliqué dans le calcul, puis diminue quand peu voire plus de consensus sont impliqués.



(a) CRS pour bestline



(b) CRS pour régression linéaire



(c) CRS pour régression polynomiale

Figure 6.11 – CRS en fonction du nombre du rayon de la cible et du nombre de points de repère dans le consensus sur le jeu de données Amazon

6.6 Aspect méthodologique pour l'utilisation des techniques

Lors de l'évaluation, trois techniques de localisation ont été testées. De plus, plusieurs conditions ont étés considérées : l'évaluation a été réalisée sur un réseau national « contrôlé » et sur un réseau mondial « non-contrôlé » et en fonction de deux variables, le rayon de la cible et le nombre de points de repère (uniquement au niveau mondial pour ce dernier point). À partir de ces éléments et des résultats obtenus, il est possible de déterminer avec quelle méthodologie utiliser les différentes techniques et dans quelles conditions elles s'exécutent le mieux.

6.6.1 Aspect contrôlé de l'environnement

Pour un utilisateur, les méthodes seront déployées très probablement dans le cadre d'un environnement non-contrôlé. Mais, si l'on regarde les résultats, ils sont meilleurs dans le cas où l'environnement est contrôlé.

Les scores plus faibles observés dans l'environnement non-contrôlé peuvent provenir du jeu de données mal nettoyé. En effet, le nettoyage du jeu de données Amazon a nécessité de visualiser certaines données afin de se rendre compte de leur incohérence. Compte tenu de la taille du jeu de données, il est possible que certains éléments affectant négativement les résultats soient toujours présents, comme des cibles situées à une autre position géographique que celle prévue. Une autre source d'erreur peut être l'estimation du score par Monte-Carlo qui ne comptabiliserait pas certains consensus.

Cependant, il est quand même probable qu'en utilisant jeu de données parfaitement nettoyé de ces éléments, les scores observés soient plus faibles au niveau national, à cause de l'aspect non-contrôlé de l'environnement, et conseiller à un utilisateur de déployer une technique au sein d'un environnement contrôlé n'est pas réaliste. Il est toutefois possible d'essayer de se rapprocher au plus des conditions contrôlées, en respectant ces critères pour les points de repère :

— Regroupés au sein d'une zone géographique précise, comme un pays, afin d'éviter des routes trop longues, qui ont plus de chances de n'être ni le plus court chemin, ni uniques.

— Homogènes en termes de configuration, pour éliminer les variations dues à des différences du délai de traitement entre deux configurations différentes. De plus, il est préférable que les points de repère soient dédiés uniquement à la mesure des métriques réseau, au moins durant le temps

où les mesures sont réalisées, ce qui permet de réduire aussi les variations du délai de traitement.

— Une bande passante garantie, même si les mesures ne nécessitent qu'une bande passante négligeable, garantir que le réseau sera disponible rapproche des conditions contrôlées.

6.6.2 Effet du nombre de points de repère

Utiliser un nombre élevé de points de repère peut mener, a priori, à de meilleurs résultats, car plus il y a de points de repère, plus le consensus sera contraint et le résultat de meilleure qualité. Mais d'après les résultats observés sur le jeu de données Amazon, pour lequel le nombre de points de repère considérés pour calculer un consensus varie, ce n'est pas le cas. En effet, un nombre trop important de points de repère pose une contrainte trop forte sur le calcul des consensus, ce qui a pour effet de ne pas trouver de consensus avec ce nombre de points de repère. Il est possible d'observer ce phénomène avec le LSS qui, malgré l'augmentation de la taille de la zone cible n'augmente pas, alors qu'il doit forcément converger vers 1. Quant au CRS, il décroit, pour les méthodes par régression, alors qu'il devrait augmenter, car les consensus sont plus restreints.

À l'inverse, ne pas utiliser assez de points de repère, revient à ne pas assez contraindre le consensus et obtenir des zones trop vastes, qui ont plus de chance de former une intersection avec la zone cible, d'où le LSS « elevé » et le CRS proche de 0.

D'après les expérimentations menées dans le cadre de cette thèse, il semble qu'utiliser 5 à 6 points de repère, bien sélectionnés, soit suffisant. Les expérimentations menées sur le jeu de données Grid'5000 utilisent 5 points de repère (et un de plus qui joue le rôle de la cible) et offrent de bons résultats. De plus avec 5 à 6 points de repère, celles menées sur le jeu de données Amazon offrent :

— Pour la bestline, un résultat acceptable en terme de LSS, avec au minimum un score de 0.5 et en termes de CRS avec un score minimum de 0.8.

— Pour les méthodes par régression, les meilleurs résultats en termes de CRS sont observés pour ce nombre de points de repère, même si le LSS reste quasiment nul. Mais même avec un seul point de repère dans le consensus, ces méthodes peinent à offrir un LSS de 0.5. De plus, au moins 3 points de repère sont requis pour une multilateration.

6.6.3 Granularité de la cible

Il est aussi possible de savoir avec quelle granularité pour la cible, ces méthodes permettent d'obtenir un résultat correct. Pour cela, on se place dans le cas optimal de fonctionnement, c'est-à-dire l'environnement contrôlé, et on peut considérer trois granularités pour la cible :

— Une ville ou d'une zone urbaine, ce qui correspond aux cercles autour de la cible jusqu'à 50 km de rayon. A ce niveau, à part pour la bestline, il est compliqué d'avoir une intersection entre le consensus et la cible. De plus, les cas où l'intersection existe elle est de mauvaise qualité car le CRS reste inférieur à 0.1. On ne peut pas considérer qu'on arrive à détecter dans quelle ville est situé la cible à moins que le consensus soit disproportionné par rapport à la cible.

— Au niveau d'une région, ce qui correspond aux cercles autour de la cible entre 50 km et 200 km de rayon. À ce niveau, à part pour la bestline, le LSS varie entre 0.4 et 0.7, une intersection est donc possible mais pas garantie. Quant au CRS varie proportionnellement à la distance. Avec les méthodes par régression, la qualité du résultat peut être considérée comme acceptable, car elles donnent un CRS de 0.6 ou 0.8 au rayon maximum considéré.

— Au niveau d'un pays, ce qui correspond aux cercles autour de la cible au-delà de 200 km de rayon. On peut encore subdiviser en trois granularités selon la taille du pays :

  — Entre 200 km et 400 km de rayon, cela correspond aux mêmes conclusions que la granularité « région », mais la probabilité d'observer une intersection est supérieure et quasiment garantie.

  — Entre 400 km et 700 km de rayon, une intersection est toujours présente. Ce niveau permet de représenter la France, qui a une superficie de 551 695 km^2, soit un cercle d'environ 419 km de rayon. Pour un cercle de cette taille, le CRS est d'environ 0.7. À ce niveau, il est possible d'avoir une intersection avec la cible et que la majorité du consensus se trouve dans cette cible, on peut considérer le résultat comme bon et l'estimation possible avec peu d'erreur.

  — À partir de 700 km de rayon et au-delà, le LSS et le CRS valent toujours 1, on peut donc être toujours sur d'avoir trouvé la bonne cible.

Ces méthodes permettent donc de trouver des cibles de manière fiable et dans un environnement contrôlé uniquement lorsqu'elles sont de la taille d'un pays. Utiliser un environnement non-contrôlé ne nous a pas permis d'avoir des estimations qui peuvent être considérées comme fiables, à part certains cas particuliers en utilisant la méthode bestline.

6.6.4 Quelle technique adopter ?

A priori, il est possible de penser que les techniques utilisant la régression (linéaire ou polynomiale) auront un comportement et des résultats différents de la bestline. Cette dernière surestimant les distance, on s'attend à ce qu'elle trouve plus souvent une intersection avec la cible. Pour les autres méthodes, qui « moyennent » la distance, on s'attend à ce qu'une intersection ne soit pas tout le temps trouvée, mais quand c'est le cas, elle est de meilleure qualité, c'est-à-dire que son CRS est plus élevé.

Ces suppositions sont vraies, d'après les conclusions des sections précédentes, la bestline trouve toujours une intersection avec la cible et la qualité de cette intersection est moindre, comparée à celle trouvée par les méthodes par régressions. Cependant, la différence dans la qualité des intersections consensus-cible entre les méthodes n'est pas si élevée en moyenne, permettant à la bestline d'être considérée comme une bonne technique malgré la surestimation des distances.

La bestline semble donc être la technique à utiliser. Néanmoins, dans certains cas, les techniques par régressions offrent une meilleure qualité de résultat. Ainsi, il pourrait être intéressant de combiner les différentes techniques, par exemple en utilisant d'abord la bestline pour obtenir une zone grossière dans laquelle les données peuvent être stockées et ensuite utiliser les méthodes par régression pour raffiner le résultat si possible.

Conclusion

1 Résumé des contributions

Avec l'augmentation constante du volume de données manipulées par les différents utilisateurs, le stockage dans le Cloud est devenu nécessaire. Ce moyen de stockage des données permet d'externaliser la gestion du stockage mais apporte une perte de contrôle sur les données. Un problème issu de cette perte de contrôle est la localisation des données. En effet, dans un contexte de stockage dans le cloud, un utilisateur n'a aucune garantie sur le lieu de stockage de ces données, bien qu'une clause concernant ce lieu puisse apparaître dans le SLA. De plus, vérifier la localisation des données n'est pas une tâche évidente et nécessite la mise en place de mécanismes par l'utilisateur ou le fournisseur concerné. Cette thèse a étudié une partie des solutions permettant cette vérification.

Pour cela, les motivations des utilisateurs pour la localisation des données ont d'abord été éclaircies. Il s'agit d'une part de préoccupations concernant la sécurité et la performance. Contrôler la position de ses données permet de les protéger en cas d'accidents, d'empêcher leur accès par des entités extérieures et de garantir leur accès au sein d'une zone administrative ainsi que réduire le délai pour y accéder et garantir une continuité d'accès. D'autre part, les préoccupations sont d'ordre légal. Plusieurs pays imposent, par leur législation, des restrictions sur le lieu de stockage de certains types de données.

Pour répondre à ces préoccupations, nous avons identifié différentes méthodes permettant la garantie de la localisation des données dans la littérature. Ce sont les méthodes utilisant des tiers de confiances et nécessitant l'implication du fournisseur dans le processus de vérification. Ces méthodes reposent sur l'utilisation de matériel sécurisé, de type TPM, ou d'applications spécifiques, qui sont mis en place au sein des infrastructures de stockage du fournisseur. Certaines de ces méthodes apportent une garantie sur la position des données, et dans ce cas l'utilisateur est en charge de vérifier que leur lieu de stockage lui convient. Pour les autres méthodes, la garantie est que les données ne seront pas déplacées en dehors des lieux fournis par l'utilisateur. Quelques failles au niveau logiciel ou matériel existent, mais elles sont facilement corrigeables ou difficilement exploitables.

Une autre famille de méthode a été identifiée, les méthodes estimant la

localisation en utilisant des points de repère. Différentes techniques existent, mais elles fonctionnent toutes de manière similaire, à la manière de techniques d'apprentissage automatique. Ceci a été mis en évidence par la classification de ces techniques, selon différents critères, que nous avons réalisée. Les résultats revendiqués par les auteurs ont aussi été reportés. Ce dernier point permet de remarquer l'absence d'unification dans la mise en valeur de ces techniques dont le fonctionnement est pourtant très proche.

Plusieurs contributions ont été faites. La première est la réalisation d'un cadre générique, qui une fois instancié, permet de concevoir les différentes étapes des techniques estimant la localisation avec des points de repère à l'aide d'une fonction d'estimation de la distance. Ce cadre permet de détailler les différentes étapes de l'apprentissage, avec la collecte de mesures, le nettoyage du jeu de données et la construction de la fonction d'estimation. Les étapes de la vérification, c'est-à-dire la collecte de mesures, le nettoyage du jeu de données et la réalisation de l'estimation donnant lieu à un résultat appelé consensus, sont aussi détaillées. Finalement, pour l'évaluation, deux scores sont proposés. Le premier est le LSS, qui donne la qualité de la technique, son taux de succès en d'autres termes. Le second est le CRS, qui donne la qualité du résultat fourni par la technique, à savoir sa précision.

Ensuite, deux jeux de données ont été collectés. Le premier, réalisé à l'aide de nœuds sur la plateforme Grid'5000, correspond à des mesures au sein d'un environnement contrôlé, c'est-à-dire un réseau privé avec des garanties d'homogénéité des configurations et une bande passante garantie. Le second, réalisé à l'aide de VM sur la plateforme AWS, correspond à des mesures au sein d'un environnement non-contrôlé, autrement dit un réseau public, internet, et sans garantie sur la bande-passante. Les deux contiennent les mêmes métriques, le RTT et le nombre de sauts. Ces jeux de données ont été nettoyés afin d'être utilisables, il s'agissait pour le premier de retirer les mesures qui ont échoué et les points de repères qui n'ont pas réalisé assez de mesures. Pour le second, en plus de ces critères de nettoyage facilement identifiable, une inspection manuelle à permis de retirer les mesures non-cohérentes. Cette collecte a été mise en place de manière « hors ligne », correspondant donc aux étapes de la collecte et du nettoyage pour l'apprentissage et la vérification.

Finalement, à l'aide des jeux de données, trois techniques ont été évaluées. Ce sont les techniques « bestline » (qui surestime les distance), régression linéaire et régression polynomiale. Différents degrés de régression polynomiale ont été testés et seul le degré 3 a été retenu compte tenu de la similarité des résultats entre les différents degrés. Les premières expérimentations sur le jeu de données issue d'Amazon ont permis de voir que dans la plupart des cas les consensus étaient rarement réalisés avec l'ensemble des points de repères. Le

calcul du consensus a donc été modifié pour prendre en compte l'ensemble des consensus possibles. Cependant, le coût de ce calcul nous a mené à utiliser une méthode de type Monte-Carlo pour le calcul des consensus.

Les résultats obtenus sont meilleurs sur le jeu de données issue de Grid'5000. C'est un résultat attendu car l'environnement été contrôlé et le jeu de données est fiable. Un utilisateur souhaitant déployer ces techniques doit donc essayer de se rapprocher le plus possible d'un environnement contrôlé pour éliminer le plus de biais dus aux mesures. De plus, d'après les résultats, un nombre de points de repères trop élevé a une influence négative sur les résultats ; limiter l'architecture à 6 points de repères semble être un bon compromis pour assurer une précision maximale. En outre, si la granularité de la cible est trop faible, il n'est pas possible d'estimer correctement le résultat. Il est impossible de détecter la cible si c'est une ville, détecter une région ou un petit pays est parfois possible et détecter les plus gros pays est toujours possible. Finalement, la technique bestline fourni des résultats convenables selon toutes les conditions et parait être la technique à privilégier. Néanmoins, les techniques par régression peuvent parfois apporter une meilleure précision des résultats, elles peuvent donc être utilisées en combinaison de la bestline pour essayer d'améliorer la précision.

2 Perspectives

Les travaux présentés dans cette thèse peuvent être poursuivis en investiguant les directions proposées ici.

— Il serait intéressant de réaliser des expérimentations avec un jeu de données national et non-contrôlé. En effet, les deux jeux de données que nous avons utilisé sont soit de type contrôlé et national, soit non-contrôlé et mondial, il manque donc une étape avec un jeu de données intermédiaire. Mettre en place cette collecte permettrait de confirmer les résultats et de mieux expliquer les résultats obtenus avec le jeu de données non-contrôlé.

— Le cadre unifié proposé permet de prendre en compte l'ensemble des techniques présentées dans le chapitre 3. Cependant, nous avons choisi de nous concentrer sur trois techniques, celles dont le fonctionnement était le plus proche. Inclure les techniques utilisant la classification [38, 39, 42] ainsi que de nouvelles techniques est une piste de recherche intéressante. De plus, il est aussi possible d'intégrer de nouvelles métriques aux techniques existantes. Dans un premier temps, il est envisageable d'inclure le nombre de sauts, car cette métrique a déjà été collectée. Cela permettrait

d'être plus exhaustif dans l'analyse et d'avoir un nouveau point de vue sur quelle technique choisir en fonctions des conditions.

— Une des limites des solutions implémentés est l'utilisation de « ping » simples pour interagir avec la cible. Dans un environnement cloud non-contrôlé, il est possible que ce mécanisme soit bloqué ou bien qu'un proxy, ne stockant pas les données, réponde à la place du serveur les stockant. Implémenter une PDP permettrait de l'éviter. Effectivement, les PDP garantissent la présence des données, car elles nécessitent d'y accéder pour fonctionner. Un accès au serveur les stockant est donc réalisé. Bien sûr, une PDP cause un « overhead » au niveau du RTT, mais comme un jeu de données sans PDP est déjà disponible, cet overhead et ses effets sur les estimations peuvent être étudiés.

— Implémenter une technique « hybride » en fonction des méthodes existantes. Par exemple, utiliser la méthode bestline pour obtenir un premier résultat, puis utiliser les points de repère les plus proches de ce premier résultat pour essayer de l'affiner avec une méthode par régression. En effet, dans certains cas les méthodes par régression donnent un résultat significativement plus précis que la bestline, mais le résultat n'est pas toujours disponible. Ainsi, dans certains cas le résultat peut être amélioré.

— Sur le long terme, améliorer la sécurité des méthodes utilisant des points de repères est envisageable. En effet, nous avons identifié différentes attaques possible sur le processus de vérification [36]. Ces attaques se présentent sous différentes formes, par exemple un fournisseur peut bloquer l'accès aux données depuis les points de repère ou bien contrefaire les RTTs. Elles interfèrent ainsi avec le processus de vérification et des contremesures méritent d'être proposées.

Publications

Article dans des conférences internationales avec actes et comités de lecture

1. **Malik Iraïn**, Zoubir Mammeri, & Jacques Jorda (2018, October). Assessment of Regression-based Techniques for Data Location Verification at Country-Level. In 2018 6th International Conference on Wireless Networks and Mobile Communications (WINCOM) (pp. 1-6). IEEE.
2. **Malik Iraïn**, Jacques Jorda, & Zoubir Mammeri (2017, December). On the Vulnerabilities of Landmark-Based Data Location Approaches : Threats, Solutions, and Challenges. In 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC) (pp. 127-134). IEEE.

Article dans des journaux de référence avec comités de lecture

1. **Malik Iraïn**, Jacques Jorda, & Zoubir Mammeri (2017). Landmark-based data location verification in the cloud : review of approaches and challenges. Journal of Cloud Computing, 6(1), 31.

Bibliographie

[1] P. Mell and T. Grance, “The NIST Definition of Cloud Computing.” [Online]. Available : https://csrc.nist.gov/publications/detail/sp/800-145/final

[2] Y. Mansouri, A. N. Toosi, and R. Buyya, “Data Storage Management in Cloud Environments : Taxonomy, Survey, and Future Directions,” vol. 50, no. 6, pp. 91 :1–91 :51. [Online]. Available : http://doi.acm.org/10.1145/3136623

[3] P. Patel, A. Ranabahu, and A. Sheth, “Service Level Agreement in Cloud Computing.” [Online]. Available : https://corescholar.libraries.wright.edu/knoesis/78

[4] Contrat de niveau de service amazon compute. [Online]. Available : https://d1.awsstatic.com/legal/amazon-ec2-sla/Amazon_EC2_Service_Level_Agreement_-_French_Translation___2018-02-12_.pdf

[5] Service Level Agreement ("SLA"). [Online]. Available : https://qbox.io/service-level-agreement

[6] Missouri tornado destroys hospital data center. [Online]. Available : https://www.datacenterdynamics.com/news/missouri-tornado-destroys-hospital-data-center/

[7] J. Eng. Four Lightning Strikes in Belgium Erase Google Customer Data. [Online]. Available : https://www.nbcnews.com/tech/internet/four-lightning-strikes-belgium-erase-google-customer-data-n412561

[8] B. Gellman and A. Soltani, “NSA infiltrates links to Yahoo, Google data centers worldwide, Snowden documents say,” 2013-10-30T05 :50-500. [Online]. Available : https://www.washingtonpost.com/world/national-security/nsa-infiltrates-links-to-yahoo-google-data-centers-worldwide-snowden-documents-say/2013/10/30/e51d661e-4166-11e3-8b74-d89d714ca4dd_story.html

[9] Text - H.R.4943 - 115th Congress (2017-2018) : CLOUD Act. [Online]. Available : https://www.congress.gov/bill/115th-congress/house-bill/4943/text

[10] Censorship of Google Sites in China | GreatFire Analyzer. [Online]. Available : https://en.greatfire.org/search/google-sites

[11] G. Pallis and A. Vakali, “Insight and Perspectives for Content Delivery Networks,” vol. 49, no. 1, pp. 101–106. [Online]. Available : http://doi.acm.org/10.1145/1107458.1107462

[12] The Digital Universe Driving Data Growth in Healthcare. [Online]. Available : https://www.emc.com/analyst-report/digital-universe-healthcare-vertical-report-ar.pdf

[13] “Code de la santé publique - Article L1111-8.”

[14] “Décret n° 2018-137 du 26 février 2018 relatif à l’hébergement de données de santé à caractère personnel.”

[15] “Règlement (ue) 2016/679 du parlement européen et du conseil du 27 avril 2016 relatif à la protection des personnes physiques à l’égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/ce (règlement général sur la protection des données) (texte présentant de l’intérêt pour l’eee).” [Online]. Available : http://data.europa.eu/eli/reg/2016/679/oj/fra

[16] Health. My Health Records Act 2012. [Online]. Available : https://www.legislation.gov.au/Details/C2018C00509/Html/Text, http://www.legislation.gov.au/Details/C2018C00509

[17] Federal Law No. 242-FZ of July 21, 2014 on Amending Some Legislative Acts of the Russian Federation in as Much as It Concerns Updating the Procedure for Personal Data Processing in Information-Telecommunication Networks (with Amendments and Additions. [Online]. Available : https://pd.rkn.gov.ru/authority/p146/p191/

[18] Notice of the People’s Bank of China on Urging Banking Financial Institutions to Do a Good Job in Protecting Personal Financial Information. [Online]. Available : http://www.lawinfochina.com/display.aspx?lib=law&id=8837&CGid=

[19] Provisions on Protecting the Personal Information of Telecommunications and Internet Users. [Online]. Available : http://www.lawinfochina.com/display.aspx?id=14971&lib=law&SearchKeyword=personal%20information&SearchCKeyword=

[20] Law of the People’s Republic of China on Guarding State Secrets. [Online]. Available : http://www.lawinfochina.com/display.aspx?lib=law&id=1191&CGid=

[21] TPM 2.0 Library Specification. [Online]. Available : https://trustedcomputinggroup.org/resource/tpm-library-specification/

[22] TPM 1.2 Main Specification. [Online]. Available : https://trustedcomputinggroup.org/resource/tpm-main-specification/

[23] C. Krauß and V. Fusenig, "Using Trusted Platform Modules for Location Assurance in Cloud Networking," in *Network and System Security*, ser. Lecture Notes in Computer Science, J. Lopez, X. Huang, and R. Sandhu, Eds. Springer Berlin Heidelberg, pp. 109–121.

[24] A. Albeshri, C. Boyd, and J. G. Nieto, "GeoProof : Proofs of Geographic Location for Cloud Computing Environment," in *Distributed Computing Systems Workshops (ICDCSW), 2012 32nd International Conference On*, pp. 506–514.

[25] G. Ateniese, R. Burns, R. Curtmola, J. Herring, O. Khan, L. Kissner, Z. Peterson, and D. Song, "Remote data checking using provable data possession," vol. 14, no. 1, p. 12. [Online]. Available : http://dl.acm.org/citation.cfm?id=1952982.1952994

[26] N. Paladi, M. Aslam, and C. Gehrmann, "Trusted Geolocation-Aware Data Placement in Infrastructure Clouds," in *2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications*, pp. 352–360.

[27] N. Santos, R. Rodrigues, K. P. Gummadi, and S. Saroiu, "Policy-sealed Data : A New Abstraction for Building Trusted Cloud Services," in *Proceedings of the 21st USENIX Conference on Security Symposium*, ser. Security'12. USENIX Association, pp. 10–10. [Online]. Available : http://dl.acm.org/citation.cfm?id=2362793.2362803

[28] L. Chen and D. B. Hoang, "Addressing Data and User Mobility Challenges in the Cloud," in *2013 IEEE Sixth International Conference on Cloud Computing*, pp. 549–556.

[29] S. Betgé-Brezetz, G. B. Kamga, M. P. Dupont, and A. Guesmi, "Privacy Control in Cloud VM File Systems," in *Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference On*, vol. 2, pp. 276–280.

[30] T. Wüchner, S. Müller, and R. Fischer, "Compliance-Preserving Cloud Storage Federation Based on Data-Driven Usage Control," in *Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference On*, vol. 2, pp. 285–288.

[31] P. Massonet, S. Naqvi, C. Ponsard, J. Latanicki, B. Rochwerger, and M. Villari, "A Monitoring and Audit Logging Architecture for Data Location Compliance in Federated Cloud Infrastructures," in *Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW), 2011 IEEE International Symposium On*, pp. 1510–1517.

[32] N. O. Tippenhauer, C. Pöpper, K. B. Rasmussen, and S. Capkun, "On the Requirements for Successful GPS Spoofing Attacks," in *Proceedings*

of the 18th ACM Conference on Computer and Communications Security, ser. CCS '11. ACM, pp. 75–86. [Online]. Available : http://doi.acm.org/10.1145/2046707.2046719

[33] P. Choi and D. K. Kim, "Design of security enhanced TPM chip against invasive physical attacks," in *2012 IEEE International Symposium on Circuits and Systems*, pp. 1787–1790.

[34] "TPM Vulnerabilities to Power Analysis and An Exposed Exploit to Bitlocker." [Online]. Available : https://theintercept.com/document/2015/03/10/tpm-vulnerabilities-power-analysis-exposed-exploit-bitlocker/

[35] A. V. Markelova, "Vulnerability of RSA Algorithm," pp. 74–78. [Online]. Available : https://elibrary.ru/item.asp?id=32865711

[36] M. Irain, J. Jorda, and Z. Mammeri, "Landmark-based data location verification in the cloud : Review of approaches and challenges," vol. 6, no. 1, p. 31. [Online]. Available : https://doi.org/10.1186/s13677-017-0095-y

[37] ——, "On the Vulnerabilities of Landmark-Based Data Location Approaches : Threats, Solutions, and Challenges," in *2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC)*, pp. 127–134.

[38] B. Biswal, S. Shetty, and T. Rogers, "Enhanced learning classifier to locate data in cloud data centres," vol. 4, no. 2, p. 141. [Online]. Available : http://www.inderscience.com/link.php?id=74248

[39] T. Ries, V. Fusenig, C. Vilbois, and T. Engel, "Verification of Data Location in Cloud Networking," in *2011 Fourth IEEE International Conference on Utility and Cloud Computing*, pp. 439–444.

[40] M. Fotouhi, A. Anand, and R. Hasan, "PLAG : Practical Landmark Allocation for Cloud Geolocation," in *Cloud Computing (CLOUD), 2015 IEEE 8th International Conference On*, pp. 1103–1106.

[41] C. Jaiswal and V. Kumar, "IGOD : Identification of geolocation of cloud datacenters," vol. 27-28, pp. 85–102. [Online]. Available : http://www.sciencedirect.com/science/article/pii/S2214212616000168

[42] K. Benson, R. Dowsley, and H. Shacham, "Do You Know Where Your Cloud Files Are ?" in *Proceedings of the 3rd ACM Workshop on Cloud Computing Security Workshop*, ser. CCSW '11. ACM, pp. 73–82. [Online]. Available : http://doi.acm.org/10.1145/2046660.2046677

[43] M. Gondree and Z. N. Peterson, "Geolocation of Data in the Cloud," in *Proceedings of the Third ACM Conference on Data and Application*

Security and Privacy, ser. CODASPY '13. ACM, pp. 25–36. [Online]. Available : http://doi.acm.org/10.1145/2435349.2435353

[44] G. J. Watson, R. Safavi-Naini, M. Alimomeni, M. E. Locasto, and S. Narayan, "LoSt : Location Based Storage," in *Proceedings of the 2012 ACM Workshop on Cloud Computing Security Workshop*, ser. CCSW '12. ACM, pp. 59–70. [Online]. Available : http://doi.acm.org/10.1145/2381913.2381926

[45] M. Eskandari, A. S. D. Oliveira, and B. Crispo, "VLOC : An Approach to Verify the Physical Location of a Virtual Machine In Cloud," in *Cloud Computing Technology and Science (CloudCom), 2014 IEEE 6th International Conference On*, pp. 86–94.

[46] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida, "Constraint-Based Geolocation of Internet Hosts," vol. 14, no. 6, pp. 1219–1232.

[47] F. Dabek, R. Cox, F. Kaashoek, and R. Morris, "Vivaldi : A Decentralized Network Coordinate System," in *Proceedings of the 2004 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, ser. SIGCOMM '04. ACM, pp. 15–26. [Online]. Available : http://doi.acm.org/10.1145/1015467.1015471

[48] Y. Chen, Y. Xiong, X. Shi, B. Deng, and X. Li, "Pharos : A Decentralized and Hierarchical Network Coordinate System for Internet Distance Prediction," in *IEEE GLOBECOM 2007 - IEEE Global Telecommunications Conference*, pp. 421–426.

[49] Y. Chen, X. Wang, X. Song, E. K. Lua, C. Shi, X. Zhao, B. Deng, and X. Li, "Phoenix : Towards an Accurate, Practical and Decentralized Network Coordinate System," in *NETWORKING 2009*, ser. Lecture Notes in Computer Science, L. Fratta, H. Schulzrinne, Y. Takahashi, and O. Spaniol, Eds. Springer Berlin Heidelberg, pp. 313–325.

[50] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson, and D. Song, "Provable Data Possession at Untrusted Stores," in *Proceedings of the 14th ACM Conference on Computer and Communications Security*, ser. CCS '07. ACM, pp. 598–609. [Online]. Available : http://doi.acm.org/10.1145/1315245.1315318

[51] B. Chun, D. Culler, T. Roscoe, A. Bavier, L. Peterson, M. Wawrzoniak, and M. Bowman, "PlanetLab : An Overlay Testbed for Broad-coverage Services," vol. 33, no. 3, pp. 3–12. [Online]. Available : http://doi.acm.org/10.1145/956993.956995

[52] A. Saied, R. E. Overill, and T. Radzik, "Detection of known and unknown DDoS attacks using Artificial Neural Networks," vol. 172,

pp. 385–393. [Online]. Available : http://www.sciencedirect.com/science/article/pii/S092523121501053X

[53] M. Shtern, B. Simmons, M. Smit, and M. Litoiu, "An architecture for overlaying private clouds on public providers," in *2012 8th International Conference on Network and Service Management (Cnsm) and 2012 Workshop on Systems Virtualiztion Management (Svm)*, pp. 371–377.

[54] R. Lee and B. Jeng, "Load-Balancing Tactics in Cloud," in *2011 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pp. 447–454.

[55] Internet Protocol. [Online]. Available : https://tools.ietf.org/html/rfc791

[56] F. Poletti, N. V. Wheeler, M. N. Petrovich, N. Baddela, E. Numkam Fokoua, J. R. Hayes, D. R. Gray, Z. Li, R. Slavík, and D. J. Richardson, "Towards high-capacity fibre-optic communications at the speed of light in vacuum," vol. 7, no. 4, pp. 279–284. [Online]. Available : https://www.nature.com/articles/nphoton.2013.45

[57] M. Irain, Z. Mammeri, and J. Jorda, "Assessment of Regression-based Techniques for Data Location Verification at Country-Level (Invited Paper)," in *2018 6th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, pp. 1–6.

[58] Grid5000. [Online]. Available : https://www.grid5000.fr

[59] Amazon Web Services (AWS) - Cloud Computing Services. [Online]. Available : https://aws.amazon.com/

[60] RENATER. [Online]. Available : https://www.renater.fr/

[61] L. Wenwei, Z. Dafang, Y. Jinmin, and X. Gaogang, "On evaluating the differences of TCP and ICMP in network measurement," vol. 30, no. 2, pp. 428–439. [Online]. Available : http://www.sciencedirect.com/science/article/pii/S0140366406003719

[62] D. S. Alves and K. Obraczka, "An Empirical Characterization of Internet Round-Trip Times," in *Proceedings of the 13th ACM Symposium on QoS and Security for Wireless and Mobile Networks*, ser. Q2SWinet '17. ACM, pp. 23–30. [Online]. Available : http://doi.acm.org/10.1145/3132114.3132123

[63] PlanetLab | An open platform for developing, deploying, and accessing planetary-scale services. [Online]. Available : https://www.planet-lab.org/

[64] *Amazon Elastic Compute Cloud (EC2)*. [Online]. Available : https://aws.amazon.com/ec2

[65] AWS Global Cloud Infrastructure. [Online]. Available : https://infrastructure.aws/

[66] G. Wang and T. S. E. Ng, "The Impact of Virtualization on Network Performance of Amazon EC2 Data Center," in *2010 Proceedings IEEE INFOCOM*, pp. 1–9.

[67] I. Bermudez, S. Traverso, M. Mellia, and M. Munafò, "Exploring the cloud from passive measurements : The Amazon AWS case," in *2013 Proceedings IEEE INFOCOM*, pp. 230–234.

[68] C. Raiciu, M. Ionescu, and D. Niculescu, "Opening Up Black Box Networks with CloudTalk," Submitted. [Online]. Available : https://www.usenix.org/conference/hotcloud12/workshop-program/presentation/raiciu

[69] Q. Jacquemart, A. B. Vitali, and G. Urvoy-Keller, "Measuring the Amazon Web Services (AWS) WAN Infrastructure," in *CoRes 2019*. [Online]. Available : https://hal.archives-ouvertes.fr/hal-02128052

[70] The Shapely User Manual — Shapely 1.7a2 documentation. [Online]. Available : https://shapely.readthedocs.io/en/latest/manual.html

[71] E. Westra, *Python Geospatial Analysis Essentials : Process, Analyze, and Display Geospatial Data Using Python Libraries and Related Tools*, ser. Community Experience Distilled. Packt Publ.